DOCUMENT RESUME

ED 218 309 TM 820 349

AUTHOR Goldstein, Harvey; Ecob, Russell
TITLE An Investigation of Models for the Estimation of Test Score Reliability Using Longitudinal Data.
INSTITUTION London Univ. (England). Inst. of Education.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [Oct 81]
GRANT NIE-G-77-0065
NOTE 143p.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Analysis of Variance; Elementary Secondary Education; *Error of Measurement; Longitudinal Studies; Mathematical Models; Path Analysis; *Scores; *Test Reliability; *Test Theory
IDENTIFIERS *Instrumental Variable Methods; Quantitative Data; *Structural Equation Models

ABSTRACT

Using data from a National Child Development Study (NCDS) in Great Britain, the applications of instrumental variable methods and structural equation models to estimating instrumental variables are presented. A subset of the longitudinal educational and home background data on children born in England, Wales and Scotland in a March week of 1958 is used. The models for quantitative variables are discussed in terms of reliability in measurement error variance, relationships and models for true score measures on two or more occasions, and identification problems in measurement error. A correction for unreliability or measurement error variance of the independent variable, instrumental variable estimation, and estimation in structural equation models are discussed. The methods used with the NCDS data are summarized. The appendices include the research project submission, a description of the NCSD, and a comparison of the classical test score model and the latent variable model. Further appendices discuss the distributions of variables, transformations, instrumental variables theory and test score reliability, LISREL applications, estimate inconsistency, missing data and a social class variable. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED218309

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

An Investigation of Models for the Estimation of Test Score Reliability using Longitudinal Data

Report of Research Project Supported by the NIE under Grant No NIE - G - 77 - 0065

By

Harvey Goldstein

and

Russell Ecob

University of London Institute of Education

Bedford Way, London WC1H 0AL

TM 820 349

An Investigation of Models for the Estimation of Test Score Reliability using Longitudinal data. Report of Research Project supported by the NIE under Grant No NIE - G - 77 - 0065

CONTENTS

Preface

Chapter 1 Models for Quantitative Variables

1.1 Models of measurement error

1.2 Relationships between true scores on two occasions

1.3 Models for true scores measured on more than two occasions

1.4 Measurement error: identification problems

Chapter 2 Methods of Estimation

2.1 Correction for unreliability or measurement error variance of the independent variable

2.2 Instrumental variable estimation

2.3 A unified approach to estimation in the just-identified case

2.4 Estimation in structual equation models

Chapter 3 Applications of Instrumental Variable Methods and Structual Equation Models to NCDS Data

Chapter 4 Measurement Error in Categorised Variables

Chapter 5 Further Research

References

Appendix 1 The research submission for the project

Appendix 2 The National Child Development Study (NCDS)

Appendix 3 The Classical test score model and the latent variable model compared

Appendix 4 The distributions of variables and transformations

Appendix 5a Theory of Instrumental variables estimation

Appendix 5b Instrumental variable methods for the estimation of test score reliability

Appendix 6 Applications of LISREL to educational data

Appendix 7. Estimating the inconsistency of instrumental variables estimates in the case of congeneric variables with correlated errors

Appendix 8. Missing data in the NCDS: the use and evaluation of a method of Beale and Little for estimating missing values

Appendix 9. Regression of attainment on social mix within social class - effect of errors in social class classification

Appendix 10. A description of two datasets used to estimate measurement error in social class.

Preface

As indicated in the grant submission for this project (Appendix 1) whilst there has recently developed a sizeable literature on statistical model building for longitudinal data, there have been few attempts to study the applicability of these models to real data. Two developments, largely since the submission, have however influenced the course of the research reported here. Firstly, the grant holder has completed his methodological research under NIE Grant No. 400-75-004 1. This covered part of the ground under part (a) of the submission (see Goldstein, 1979) and drew attention to the need to compare instrumental variable estimators using different choices of instrumental variables with regard to their consistency. Secondly, more progress has been made in the use of Structural Equation models in longitudinal data, particularly by Karl Joreskog and his co-workers in a project on "Statistical methods for the analysis of longitudinal data", and the computing difficulties have been substantially alleviated with the availability of the LISREL program (Joreskog and Sorbom, 1978), now into its 5th version. This suggested the use of these procedures on the NCDS data.

As a result of the first development, a substantial part of this project has been devoted to the examination of the method of instrumental variables estimation. Estimates obtained using different variables as instrumental variables are compared in the light of theoretically derived hypotheses about their relative values in the regression of 16 year on 11 year scores, separately for tests of reading and mathematics (see Appendices 5a, 5b).

Structural Equation models were applied in an exploratory sense to the regression of 11 year on 7 year reading attainment and in a confirmatory sense to the relationship between reading attainments over the three ages 7, 11 and 16, and the parameter estimates compared with those obtained using the instrumental variables method (see Appendix 6). In addition, a re-analysis is given of an application of Structural Equations models to reliability estimation on longitudinal data showing the dependence of the estimates on the particular final models used (see Appendix 6). Expressions for the inconsistency of Instrumental Variables estimates in terms of the correlations of the errors of measurement of the variables involved are given and Structural Equations methods are used to obtain estimates of these correlations (see Appendix 7).

Preliminary to any analysis of such a dataset as the National Child Development Study (described in Appendix 2) it is necessary to check on the distributional characteristics of the relevant variables and if necessary to carry out transformations. A discussion is given (see Appendix 4) of possible transformations of variables and the conditions for their use. Here the

possible conflict between transformations which give linearity of relationships and those which give marginal normality is examined when the data do not posses the property of multivariate normality. In addition when many variables are used simultaneously in an analysis, such as when using a number of instrumental variables, background variables or multiple indicators of a latent variable, the problem of partial non-response is highlighted where one or more of the relevant variables are missing for a particular case. A method of interpolation of partial non-response due to Beale and Little (1976) is examined on the NCDS data (see Appendix 8) and a comparison is made of estimates obtained by this method which uses the information from partial non-respondents, with estimates obtained when all such cases are deleted.

Some thought has been given to the analysis of models of measurement error in categorical data in particular for obtaining measures of change in true social class between two ages by correcting the observed social mobility matrix for error in social class as suggested in the discussion in Goldstein (1979a). This work is not reported here since we have been unable as yet to solve the computing problems involved in obtaining reasonable estimates of the conditional probabilities relating the true social class probabilities at different ages. However, estimates of measurement error of social class obtained from three data sets (see Appendix 10) were used in a model of the regression of attainment on a measure of social class mix correcting this measure for error in social class (see Appendix 9). This model was tested on data from a large literacy survey and the correction for measurement error was shown to alter the conclusions substantially. Other points in the research submission not examined in detail are the study of log-linear models and of scoring methods for categorical data. On the log-linear models, initial work failed to give promising results. This is not to deny the potential of these methods and the comparisons suggested in the research submission are still considered valuable. The scoring methods for categorical data were not examined in detail, though an investigation was made of the effect on the parameters of the structural equation models of alternative scoring methods for the teacher ratings (see Appendix 6).

The report has been divided, for convenience, into five chapters and ten appendices, with individual responsibility for the latter being indicated where appropriate.

Russell Ecob
Harvey Goldstein

October 1981

REFERENCES

1. GOLDSTEIN, H (1979) Some Models for Analysing Longitudinal Data on Educational Research. J.R.Statistical Soc. 142 407-442

2. JORESKOG, K G & SORBOM, D (1978) LISREL IV: Users Guide

ACKNOWLEDGEMENTS

We wish to thank the following people who contributed to the Project: Ian Plewis, Steven Simpson and Dougal Hutchison who provided useful comments from time to time. Shirley Freeman and Kay Pilpel who provided valuable help with the typing.

1. Models for Measurement of quantitative variables

1.1 Models of measurement error

We briefly describe classical test score theory and latent variable theory, give a definition of reliability and show its relation to measurement error variance.

Let x_{ijk} be a measurement of individual i on a test item j at occasion k, then the classical test theory model is that

$$x_{ijk} = T_{ij} + u_{ijk} \qquad (1)$$

where u_{ijk} is the measurement error and T_{ij} is the true value or true score. The error is a random variable defined as having zero expectation over replications and zero correlation with the true score. The variance of the measurement error is σ^2_{ujk}.

Though sometimes described as a latent variable model this model is essentially different in that expectations are taken over replications within individuals as opposed to over individuals in the latent variable model. This latter model is contrasted with the classical test theory model in Appendix 3, where a formal axiomatic basis for the two models are given and it is shown that an additional axiom is required for the latent variable model in order that it can be used in the classical test theory context, and traditional reliability estimates used. This extra axiom is that the covariance of errors of any two individuals over replications is zero and is called the axiom of experimental independence between persons. Conditions under which this axiom may not hold are given in Appendix 3 as well as references to the relaxation of this axiom. This axiom will be assumed to hold for the remainder of this report where latent variables and classical test theory are used interchangeably. We will use the term true score in all contexts.

For a given occasion and item or test, x, the reliability (R) of x is given by (we omit the individual subscripts from now on)

$$R = \sigma^2_T / \sigma^2_x \qquad (2)$$

where σ^2_T is the variance of true scores over individuals and thus

$$\sigma^2_T = \sigma^2_x - \sigma^2_u \qquad (3)$$

For a test composed of many items an assumption of local independence between items is necessary for the use of both classical test theory and latent variable theory. This is described also in Appendix 3 and it is shown

that it is required in the formulae for the traditional reliability estimates which make use of the relations between the items in a test.

1.2 Relationships between true scores on two occasions

One possibility for modelling the relation between variables is to standardise the distributions at each occasion thus assuming a common scale. When this is done and the crude difference between the variables is considered as a measure of change we have the unconditional model for change over time. This model has deficiences when comparing changes between 2 or more subgroups of the population (formed by dividing the population say in terms of sex of social class) as the variation within each of the groups may not then be constant across occasions. In addition the error variances on the two occasions are unknown and possibly unequal for different subgroups, so that the subgroup variances of the true scores on the separate occasions are unknown.

In the remainder of this report we consider the conditional regression model for true scores,

$$T_2 = \alpha + \beta T_1 + e \qquad (4)$$

The rationale for choosing this model is discussed in detail by Goldstein (1979). Briefly, inferences drawn from this model are robust against non-linear scale transformations and the model implicitly incorporates the time asymetry present in the real world. Neither of these properties is shared by the simple change model.

1.3 Models for true scores measured on more than two occasions

We consider measurement at three separate occasions and the generalisation to more than three is straightforward. Linear regression equations relating the true scores for a given attainment over the three occasions can be written as

$$T_2 = \alpha_1 + \beta_1 T_1 + e_1 \qquad (5)$$

$$T_3 = \alpha_2 + \beta_2 T_1 + \gamma_2 T_2 + e_2 \qquad (6)$$

A system of equations of this form is called a recursive system as for each new equation in the system a variable is introduced which is not present in any previous equation in the system. The identification of this system, or the existence of unique parameter estimates is discussed by Johnston (1972, p 365). It requires in particular that the correlation between errors

e_1, e_2 is known.

A zero correlation between these errors is necessary in order that the Ordinary Least Squares (OLS) estimates when applied to each equation separately gives efficient estimates and for each equation a zero correlation between these errors and the independent variables is necessary for consistent estimates.

Neither of these conditions will hold if the equations are mis-specified, for instance by the exclusion of a relevant variable, or for example by the exclusion of a quadratic or higher order term in one of the independent variables in the equation.

For the NCDS data both for reading and mathematics attainment, almost linear relationships between observed scores at each pair of ages 7, 11 and 16 can be obtained by suitable transformations of test scores which also produce a ratio of maximum to minimum variance around the regression line not exceeding 2.0. This is achieved for both attainments by empirical transformations which give standard normal distributions for the 11 and 16 year scores, the 7 year reading test value for reading being transformed to give a linear relationship with the 11 year scores. The mathematics 7 year raw score is roughly normally distributed and linearly related to the 11 year score without transformation. The transformation of the 7 year reading score is required because of the strong ceiling effect on this test (22% of the observations were in the top 2 out of 30 values) and though rendered more normal by this transformation, the skewness and kurtosis remain high, -1.1 and 3.3 respectively.

Thus mis-specification in the above models will be due only to omitted variables rather than non-linearities. One relevant omitted variable is Social Class and including this variable as a set of dummy variables (w_j). Equations (5) and (6) become

$$T_2 = \alpha_1 + \beta_1 T_1 + \sum_j \delta_{1j} w_j + e_1 \qquad (7)$$

$$T_3 = \alpha_2 + \beta_2 T_1 + \gamma_2 T_2 + \sum_j \delta_{2j} w_j + e_2 \qquad (8).$$

This model still does not accommodate changes in Social Class between the occasions and this is either accomplished by including the values at each occasion in equation (8) or by including the value at the 2nd occasion and the change between the 1st and 2nd occasion, the latter giving an easier interpretation as well as more precise parameter estimates, since social class is highly associated at occasions 1 and 2.

The model can be further extended to the case where two dependent variables, for example mathematics and reading attainment, are related across occasions. This case has been considered in detail by Goldstein (1979) and we shall not pursue it further.

1.4 Measurement error: Identification Problems

Consider first the case of one independent variable and the relationship between the true scores given in equation (4).

We now add the measurement equation from classical test score theory (equation 1)

$$x_1 = T_1 + u_1 \tag{9}$$

$$x_2 = T_2 + u_2 \tag{10}$$

where Var $(T_1) = \sigma^2$, Var $(u_i) = \sigma_{ui}^2$, and assume that under the classical test theory axioms the covariances between the measurement errors u_1, u_2 and between the measurement errors, u_1, u_2 and the disturbance term in (4), are all zero.

If we assume in addition that the variables T_1, T_2 and the errors u_1, u_2, e are normally distributed then all the distributional information is contained in the first two moments as the observed variables are also normally distributed.

We have

$$\begin{aligned} E(x_1) &= \mu_1 \text{ say} \\ E(x_2) &= \alpha + \beta\mu_1 \\ \text{Var } (x_1) &= \sigma^2 + \sigma^2_{u_1} \\ \text{Var } (x_2) &= \beta^2\sigma^2 + \sigma_v^2 \\ \text{Cov } (x_1\ x_2) &= \beta\sigma^2 \end{aligned} \tag{11}$$

where $\sigma_v^2 = \sigma_{u_2}^2 + \sigma_e^2$

which expresses the 7 unknown parameters in terms of the 5 observed means, variances and covariances. Note, however that σ_v^2 is sufficient for inferences about β, so that we will consider only the estimation of the six unknowns, $\alpha, \beta, \mu_1, \sigma^2$, $\sigma^2_{u_1}$, σ^2_v. These equations may also be obtained from the likelihood function of the observations. By examination of these equations it can be seen that once β is determined, σ and μ_1 can be found by substitution. To solve the remaining equations we need a further restriction. One possible restriction is $\sigma_{u_1}^2 = k\sigma^2$ which corresponds to the use of a value R_1 of the reliability of x_1 viz, $k = (1 - R_1)/R_1$. Other possible restrictions are $\sigma_{u_1}^2$ known, or the ratio $\sigma^2_{u_1}/\sigma^2_v$ known. The restriction $\sigma^2_e = 0$ will sometimes hold in the physical sciences and is the case dealt with by Madansky (1969).

The equations (11) set may also be identified by using extra information from other variables. These may be either replicate observations on x_1

or other, correlated, variables known as 'instrumental' variables.

With replicate observations where their measurement errors are uncorrelated, this provides an estimate of σ^2 and hence R leading to identification.

The instrumental variable, Z, is assumed to have zero covaria with the errors u_2, u_1 and e. We then have $\text{cov}(x_2 Z) = \beta \text{ cov}(x_1 Z)$ and we obtain an estimate of β which allows the other 3 parameters to be determined uniquely.

We may also ask what are the general conditions under which this model without extra information is not identified. It turns out (Riersol, 1950) that the only conditions under which identification does not hold is when T_1, T_2 are normally distributed or are constant; the lack of identification resulting from the absence of information about the parameters from moments higher than the second which are all zero in the normal distribution case. A generalisation of this condition to the multiple regression case is that the parameter vector β is identified if and only if there exists no linear combination of the vector T which is normally distributed(Aigner,Kapteyn and Wansbeek, 1981). The simple model is identified even when T_1 is normally distributed if neither the distributions of $\sigma_{u_1}^2$ or σ_v^2 have a normal distribution (Riersol, 1950).

Where u_1, u_2 and T have non-zero covariances then independent estimates of these values are required for identification.

2. Methods of Estimation

2.1 Correction for unreliability or measurement error variance of the independent variable

For the one independent variable case, we have from (4)

$$x_2 = \alpha + \beta x_1 + (e - \beta u_1 + u_2)$$

and it can be seen that, due to the presence of the term u_1 whose correlation with x_1 is (1-R), the error term is now negatively correlated with x_1.

The ordinary least squares (OLS) estimator, $\hat{\beta}_{OLS}$ of β then has expectation in the limit as the sample size tends to infinity,

$$E(\hat{\beta}_{OLS}) = \beta \left\{ \frac{\sigma_u^2}{\sigma_{x_1}^2 + \sigma_u^2} \right\} = \beta R \qquad (12)$$

and a consistent estimator for β is thus obtained by

$$\hat{\beta}^*_{OLS} = \hat{\beta}/R \qquad ( )$$

In finite samples of size n the expectation of $\hat{\beta}^*_{OLS}$ has been shown by Richardson and Wu (1970) to be $E(\hat{\beta}^*_{OLS}) = \beta \{ 1 + \frac{2R(1-R)}{n} + O(\frac{1}{n^2}) \}$. This gives a bias of less than 1 in 10,000 for the present data.

In the NCDS data then, where the reliability (R) is not known precisely but an unbiased estimate of R, distributed independently of $\hat{\beta}_{OLS}$, is available, then $\hat{\beta}^*_{OLS}$ is almost unbiased. The measurement error of the reliability estimate itself will inflate the variance of $\hat{\beta}^*_{OLS}$ and this can be taken into account (Fuller and Hidiroglou (1978)).

A number of similar methods are available for the multiple regression case to take into account known or estimated variances and covariances of measurement errors (Hidiroglou, Fuller, Hickman, 1979).

Estimation procedures have been described for the following cases

a) when the reliability of each independent variable is either known or estimated, when the reliability of the dependent variable may or may not be known (Fuller and Hidiroglou, 1978).
b) for general Σ_{uu}, Σ_{uw}, σ_w^2 being known or for which estimates are available, σ_e^2 being unknown but positive (Fuller, 1980).
c) when Σ_{uu}, Σ_{uw} only are known and where σ_e^2 and σ_w^2 are unknown but positive (Hidiroglou, Fuller and Hickman, 1979) and
d) for the above case where Σ_{uw} is zero, Σ_{uu} not being assumed

diagonal (Fuller, 1980), the specialisation to diagonal Σ_{uu} being given by (Warren et al, 1974).

All these results apply more generally for multiple dependent variables. The estimation methods are all based on least squares and make no assumption about the distribution of the independent variables.

2.2 Instrumental Variable estimation

A drawback to these methods is the need to make the assumption that Σ_{vw} is known or in particular is zero or can be estimated. The Instrumental Variable methods which use extraneous information provide consistent estimates of $\hat{\beta}$ and $\text{Cov}(\hat{\beta})$ even when Σ_{vw} is unknown, provided the instrumental variable used in conjunction with a particular independent variable is uncorrelated with the error in both the independent and dependent variables and also with the equation disturbance term.

This method was used by Goldstein (1979) to correct the measurement error of 7 year scores of reading and mathematics using teacher ratings at the same age, since no good estimate of the reliability of the test was available. Ecob and Goldstein (1981) examined the suitability of instrumental variable estimation for estimation of change in reading and mathematics between the ages of 11 and 16 in the same study by comparing estimates using different instrumental variables and after having formulated hypotheses as to their likely values. This paper is reproduced as Appendix 5b, the theory of the method being given in Appendix 5a.

As the OLS estimator consistently estimates β/R, the instrumental variable estimator will also give a consistent estimate of the reliability (R) of the independent variables by dividing the OLS estimator by the instrumental variable estimator. An expression for the asymptotic Variance-Covariance estimate of the vector of regression coefficients, measurement errors of independent variables and disturbances is given in Kapteyn and Wansbeck (1978) which enables the standard errors of the reliability estimates to be obtained.

2.3 A unified approach to estimation in the just-identified case

Kapteyn and Wansbeck (1978) present an estimator for the multiple regression situation of which includes the estimator in a) and b) above as special cases.

The consistent adjusted least squares (CALS) estimator is of the form

$$\hat{\beta} = (X'X)\ (X'X-nC)^{-1} b_{OLS}$$

where C is the variance covariance matrix of the errors, u in X and b_{OLS} is the ordinary least squares regression estimator. As C is not generally known an identifying restriction is made which is either exact, in general $F(\hat{\beta}, \sigma_e^2, C) = 0$ or stochastic, $F(\hat{\beta}, \sigma_e^2, C(\lambda)) = 0$ where λ is an unknown vector of random variables.

2.4 Estimation in Structural equation models

The two principal programs available COSAN (McDonald, 1980) and LISREL (Joreskog and Sorbom, 1981) both now offer a variety of estimation methods (least squares, generalised least squares and maximum likelihood). (See also Bentler & Weeks, 1980).

The option of generalised least squares estimation in LISREL V (Joreskog and Sorbom, 1981) allows the modelling of data which are not of multivariate normal form, the maximum likelihood estimates having unknown distributional properties. The program used in the application of structural equation modelling in Appendix 6, LISREL IV, uses maximum likelihood estimation methods and some investigation of the effect of the non normal distribution on the parameter estimates is made.

3. A Summary of the results of Instrumental Variable estimation and Structural equation modelling applied to the NCDS Data.

We here summarise the approach used and conclusions reached by Ecob and Goldstein (1981) using instrumental variables. A number of possible variables were examined separately as possible choices of instrumental variables in the estimation of regression of attainment at 16 years on attainment at 11 years in reading and mathematics separately. These included teacher ratings at ages 7, 11 and 16 of a variety of attainments and skills and also the social class of the father when the child was at each of these ages. Then a number of hypotheses were set up, motivated by theoretical expectations regarding the relationship between particular instrumental variables and the errors of measurement in the independent and dependent variables separately and the disturbance term of the regression equation. These related for the teacher ratings, to whether they measured the same or different attainment and whether they were measured on the same occasion as the independent or dependent variables. The results suggested that teacher ratings on the same attainment as that tested when taken at the same time as the tests were positively correlated with test score error, and that this correlation was lower when the teacher rating was of a different attainment from that tested but still persisted when the rating was taken at a different time from the test. However, teacher ratings were uncorrelated with the disturbance terms. In contrast, social class was correlated with disturbance terms though not with test score error. Whilst none of the instrumental variables exactly satisfied the conditions for a consistent estimate, the correlations with test score measurement errors of the teacher ratings worked in opposite directions for the dependent and independent variables. Excluding ratings taken at the same occasion as the dependent variable and also social class gave estimates of the regression coefficient within a reasonably narrow range. (0.94 to 0.99 for reading attainment and 0.84 and 0.92 for mathematics attainment) which was of the same order as the standard error (0.13 to 0.18).

The estimated standard errors using suitable instrumental variables individually is shown to be less than was obtained by the split half method used by Goldstein (1979) on 300 cases though not as low as obtainable by the split half method applied to the whole data. The reliabilities of reading and mathematics attainment were also examined separately in different social classes by this method and different values for estimates of both reliabilities and of the variance of measurement error were found. These allowed estimates of the true correlation between attainments at each age within social class to be made.

The structural equation modelling approach is applied to the NCDS data in Appendix 6 and we briefly summarise here the procedures used and conclusions reached. Though not always clear cut there is a distinction to be made between exploratory analyses in which the model is systematically extended to involve larger numbers of parameters in order to provide a better fit to the data and confirmatory analyses where restrictions are made to the model and accepted according to tests of fit.

The exploratory analyses were used in an investigation of reading attainment at ages 7 and 11. The change in reading attainment between these ages was examined using the conditional regression relation of equation (4). The analyses suggested that the addition of either of two extra indicators had little effect on the parameter estimates.

The relationship between reading attainment at the three ages 7, 11 and 16 was then examined. A substantial improvement in fit was found by assuming a test specific factor for the reading tests at each age and this was found to load particularly highly on the reading tests at 11 and 16 (the same reading test was used on these occasions). The addition of a test specific factor for teacher ratings further improved the fit of the model.

The estimates of the structural relationship parameter were compared with the instrumental variable estimates and broad agreement was found.

4. Measurement Error in categorised variables

The estimation procedures for models with qualitative variables subject to measurement error assume constancy of measurement error distributions, independent of true values. Where the measurement error is in discrete or categorised variables, however, this distribution will not generally be independent of the true value or category. Thus, the probablity of misclassifying an observation will in general depend on the true underlying category to which it belongs.

In Appendix 8, a simple model is proposed for analysing measurement errors in social class, assuming just two categories and known misclassification probabilities. The results show that quite large adjustments to model parameters are obtained when estimating true score coefficients and this suggests that there is a need systematically to develop methods for dealing with such data.

5. Further Research

The project has shown how a large longitudinal data set can be used empirically to provide estimates of measurement error variance. Two particular areas of further research have also been identified viz.

1. The extension of structural equation models to handle partial information (i.e. sample estimates) about measurement error variables.

2. The development and the empirical testing of models for measurement error in categorical data.

It is our view also, that the present project has demonstrated the need for empirically based data analysis to study the assumptions of the various models of measurement error. While we see a need for yet more theoretical development, there is a danger that this could outrun the ability of existing data to discriminate between alternative model assumptions. In particular, the NCDS data and other similar large data sets should be exploited fully in the development of new techniques.

REFERENCES

Most references for instrumental variables estimation, structural equation modelling and effects of social mix are contained in the relevant appendices and will be omitted from this reference list. This list, however includes references for Appendices 2, 3, 4, 8, 10.

AIGNER, D J, KAPTEYN, A & WANSBEEK, T J (1981) Latent Variables in Economics and elsewhere In Intrilligator: Handbook of Econometrics (forthcoming)

BEALE, E M L & LITTLE, R A (1975) Missing Values in Multivariate Analysis. Journal of the Royal Statistical Society, B, 37, 129-145

BENTLER, P M & WEEKS, D G (1980) Linear Structural Equations with Latent Variables. Psychometrika, 45, 289-305

BOX, G E P & COX, D R (1964) An Analysis of Transformations. Journal of the Royal Statistical Society B, 26, 211-252

ECOB, R & GOLDSTEIN, H (1981) Instrumental Variable Methods for the Estimation of Test Score Reliability Submitted for Journal publication

FOGELMAN, K (1976) Britain's Sixteen Year Olds National Children's Bureau, London

FOX, T & ALT, J (1976) The Reliability of Occupational Coding. Paper presented to the Seminar on Longitudinal Studies New College, Cambridge, March 1976

FULLER, W A & HIDROGLOU, M A (1978) Regression Estimation after Correction for Attenuation. Journal of the American Statistical Association 76, 99-104

FULLER, W A (1980) Properties of some estimators for the errors-in-variables model. The Annals of Statistics, 8, 407-422

GEARY, R C (1943) Relations between statistics: the general and the sampling problem when the samples are large Proc. R. Irish Academic A, 49, 177

GOLDSTEIN, H (1979) Some Models for analysing data on Educational Attainment (including Discussion) Journal of the Royal Statistical Society A, 142, 407-422

GUTTMAN, L (1945) A basis for analysing test-retest reliability Psychometrika, 10, 255-282

GUTTMAN, L (1953) Reliability formulas that do not assume experimental independence. Psychometrika, 18, 225-239

GUTTMAN, L (1969) Review of Lord & Novick: Statistical Theories of Mental Test Scores. Psychometrica, 34, 398-404

HEALY, M J R & GOLDSTEIN, H (1976) An approach to the Scaling of Categorised Attributes. Biometrika, 63, 219-229

HOPE, K, GRAHAM, S & SCHWARZ, J R (1974) Uncovering the Pattern of Social Stratification: a two year test-retest enquiry. Internal Report, Nuffield College, Oxford

JORESKOG, K G & SORBOM, D (1981) LISREL V Users Guide (forthcoming)

HIDIROGLOU, M A, FULLER, A HICKMAN, R P (1979) SUPERCARP 5th Edition Statistical Laboratory, Iowa State University

KELDERMAN, H (1981) LISREL Models for Inequality Constraints in Factor and Regression Analysis. Paper read to the Seminar on structural equation modelling with particular reference to LISREL, held at the Institute of Education, London University, September 1981

KAPTEYN, A & WANSBEEK, T J (1978) Errors in Variables: Consistent and adjusted leastsquares estimation. In: Multivariate Analysis V (1980) Ed: Krishnaiah

KEMPF, W (1977) Dynamic Models for the Measurement of "Traits" in social behaviour. In Kempf, W & Repp, B (Eds) Mathematical Models for Social Psychology, Huber; Bern, Wiley, New York.

KENDALL, M G & STUART, A (1973) The Advanced Theory of Statistics, Vol 2, Third Edition, Griffin, London

LORD, F M & NOVICK, M R (1968) Statistical Theories of Mental Test Scores Addison-Wesley, Reading, Massachusetts

MARINI, R, OLSEN & RUBIN, D (1979) Maximum Likelihood Estimation in Panel Studies with Missing Data In K F Schuessler (Ed) Sociology Methodology, 1980, Jossey-Bass, San Francisco

McDONALD, R M (1981) The Dimensionality of Tests and Items British Journal of Mathematical and Statistical Psychology, 34, 100-117

RICHARDSON, D H & WU, D M (1970)
Least Squares and Grouping method estimation in the errors in variables model.
Journal of the American Statistical Association, 65, 724-748.

RIERSOL, O (1950)
Identifiability of a linear relation between variables which are subject to error.
Econometrika, 18, 375-389

SCOTT, E L (1950)
Note on consistent estimates of the linear structural relation between two variables.
Annals of Mathematical Statistics, 21, 284-288.

WARREN, R D, WHITE, J K & FULLER, W A (1974)
An errors-in-variables analysis of managerial role performance.
Journal of the American Statistical Association, 69, 886-893

APPENDIX I RESEARCH SUBMISSION

Introduction

Since the bringing together of the first broad collection of papers dealing with the problems of longitudinal studies (Harris, 1963), research workers, especially in the child development field, have become extremely interested in designing and analysing such studies. It has been recognised by such workers that certain questions of interest can be answered only with longitudinal data and hence there has been a practical stimulus to the development of appropriate methodology, especially that concerned with statistical model building. Much of this model building has developed from the original covariance structures model of Joreskog (1970), and there is now a sizeable literature dealing with alternative models for analysis; a useful short bibliography is given by Joreskog and Sorbom (1976).

Along with these developments, however, there seem to have been few attempts to study the applicability of different specialisations of the models to real data. Because of the need to incorporate parameters, especially measurement errors and high order time lags, these models tend to be overparameterised. Thus, in any practical application particular parameter values or relations between parameters need to be specified. There is a similar problem with more traditional techniques such as factor analysis, and experience with their application to real data suggests that the problems are not easy to resolve.

It appears to the applicant, therefore, that a useful contribution to the subject at its present stage of development would be the testing of some of the assumptions in these models with a view to obtaining specialisations which come as close as possible to realistic descriptions of actual data. Two broad approaches are available in tackling this question. First, one may attempt to simulate realistic situations and hence compare the performance of alternative models. Although useful, this approach would be not only very time consuming, but would lose much of its usefulness

unless the simulate structures were in fact known to be realistic. For this reason it seems logically to come after a second approach has been tried, and with which this application is largely concerned. This second approach involves the application, testing and further development of mathematical and statistical models for the analysis of longitudinal educational and social data, using data obtained from an extensive and representative sample of individuals. The data set it is proposed to use, known as the National Child Development Study, is briefly described in the next section, following which the specific aims of the project are detailed.

The National Child Development Study

The data set which will be used in the investigation consist of measurements made on the total cohort of children born in Britain during 3rd-9th March, 1958. These 17,000 babies were the subject of a large survey at birth, and at the ages of 7, 11 and 16 years. At the three latter ages, a large amount of educational data were obtained as well as social, physical and medical data. At the age of 16, about 87% of the survivors still living in Britain provided information, some 14,000. Preliminary investigations (Goldstein, 1976) suggest that no serious response bias exists for the basic educational variables to be used by the project.

The size and representativeness of this sample of children is unrivalled. It can be used to make valid inferences about the development of the child population of Britain from birth until the last year of compulsory schooling. It is also a large enough sample to study satisfactorily the performance of children when test scores are categorised into narrow intervals. It has data covering a very wide range of child development, thus allowing relationships between different aspects of development to be studied.

Because of its size, this sample can be expected to give fine discriminations between alternative models. The distributional forms of error terms can be studied in detail, as can various assumptions of independence between such terms. Furthermore, the size of the sample allows one to appeal to the idea of 'consistency' when making parameter estimates and carrying out significance tests, so avoiding some of the difficult problems associated with the usual maximum likelihood and related estimation procedures.

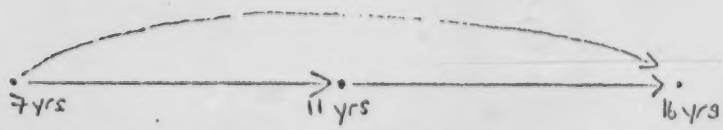
The applicant has been associated with the National Child Development Study for over 10 years, and is at present engaged on methodological research using these data under an NIE contract (No. 400-76-0041). The National Children's Bureau has agreed to make available a data tape containing the variables relevant to the proposed project, for the purpose of carrying out the work. It has agreed to this on the grounds that because of past involvement, the applicant has the necessary understanding and experience of the data to pursue independent methodological research with it.

Outline of the Project

a) The data to be used are those collected at the three ages of 7, 11 and 16 years. The basic model can be written as follows and is illustrated by the accompanying path diagram.

$$y_i = \alpha_1 + \beta_1 x_i + \varepsilon_{1i}$$

$$z_i = \alpha_2 + \beta_2 x_i + \gamma y_i + \varepsilon_{2i}$$



Where x_i is the 7 year measurement on a child, y_i is the 11 year measurement, and z_i the 16 year measurement, with the usual meanings attached to the other symbols. The measurements

used, in turn, are tests of reading and mathematic attainments.

To begin with, it is clear that the above 'non recursive' system of equations involves assumptions of linearity and additivity, and these can readily be tested with the availability of such a large sample. Transformations of the data will be studied, designed to satisfy these assumptions. In addition, the distributions of the error terms will be examined, especially with regard to normality and homoscedasticity assumptions, and mutual independence of error terms.

The first major problem with this system arises when one wishes to recognise the 'fallibility' of the measurements used. That is, if one wishes to make inferences about 'true' underlying attainments as opposed to inferences about relationships between observed scores, then the unobservable 'measurement error' of the tests used must be incorporated into the system. It is well known that the parameter estimates in the above equations are inconsistent estimates of the parameters in the corresponding equations relating the true attainments. In order to provide 'good' estimates of these latter parameters, which are at least consistent, further information must be provided. This may for example be in the form of additional equations involving 'instrumental' variables, or in the form of independent estimates of the variances and covariances of the measurement errors. The first approach involves further assumptions about independence of error terms, and these will be studied. The second approach will yield consistent estimates, but depends on either known population values of the variances and covariances or good stochastic estimates. The latter are available for the measurements used, and results with the two approaches will be compared, thus providing further checks on particular assumptions. An early analysis along these lines is described by Fogleman and Goldstein (1976).

On top of this basic model, explanatory variables will be introduced at each age and their relationship with the dependent variables examined. For some of these variables, such as family size, there is interest also in the effects of changes in the variable between ages, and the most useful method of incorporating such change variables into the model will be investigated. Finally, a 'bivariate' model will be examined where both reading and mathematics scores at each age are incorporated into the model. With these more complex systems of equations, the large sample size will again permit careful examination of alternative model assumptions.

b) Most of the methodological literature on longitudinal analyses deals with continuous measurement data. Some work, using log-linear models, has also been done for discrete or categorised data (Goodman, 1973). Of considerable interest, however, is the relationship between the two approaches. For example, to some extent assessment and progress through an educational system is based upon categorisations of essentially continuous underlying abilities, and some of the consequences of this will be considered by comparing the relationships across time of educational categories imposed by teachers (as expressed in ratings) and the relationships between continuous variables described above. The comparison will also be carried out using categorisations of the continuous variable measurements themselves. The size of the sample will allow useful comparisons to be made between these approaches, and attention will also be paid to the little discussed problem of measurement error in categorical data.

c) Many educational data consist of item rating scales, for example for behaviour on academic motivation. Where a set of item ratings is intended to reflect an underlying attribute, for example behaviour towards a teacher, it is convenient to allot scores to the item categories to give an overall score for each child. These overall scores can then be treate as pseudo-continuous in subsequent analyses. Various procedures can be used to estimate the scores, both across-sectionally and taking account of the longitudinal nature of the data. A detailed discussion is given by Healy and Goldstein (1976) and the project will extend their results in two directions.

First, by applying the techniques to the National Child Development Study data on behaviour and academic motivation. In particular, scales will be related across ages in order to devise scoring systems which agree as closely as possible at each age, and to compare these with those derived separately at each age. The techniques will also be used in order to search for meaningful sub-scales. Secondly, to extend the techniques by looking at the possiblity of alternative 'constraint' systems suggested by the data, to look at the possibility of 'rotating' estimated vectors, and to study the distributional properties of the scales. As before, the extensiveness of the data will enable a proper assessment to be made of the practical usefulness of these different scoring systems.

References

1. Fogelman, K.R. and Goldstein, H. (1976) Social Factors Associated with changes in Educational Attainment Between 7 and 11 Years of Age. Educational Studies 2, 95-109

2. Goldstein, H. (1976) A Study of Response Rates of Sixteen Year Olds in the National Child Development Study. In Britain's Sixteen Year Olds. Ed. K. Fogelman, National Children's Bureau, London.

3. Goodman, L.A. (1973) Causal Analysis of Data from Panel Studies and Other kinds of Surveys. Amer. J. of Sociol. 78, 1135-1191.

4. Harris, L.W. (1963) (Ed.) Problems in Measuring Change. Univ. of Wisconsin Press, Madison.

5. Healy, M.J.R. and Goldstein, H. (1976) An Approach to the Scaling of Categorised Attributes. Biometrika, 63, 219-229

6. Joreskog, K.G. (1970) A General Method for the Analysis of Covariance Structures. Biometrika, 57, 239-251.

7. Joreskog, K.G. and Sorbom, D. (1976) Statistical Models and Methods for Analysis of Longitudinal Data. Research Report 76-1 Department of Statistics, Univ. of Uppsala.

APPENDIX 2

THE NATIONAL CHILD DEVELOPMENT STUDY

The National Child Development Study (hereafter called NCDS) consists of a cohort of around 17000 children comprising all births in England, Wales and Scotland in the week 3rd - 9th March 1958. The initial purpose of the study was to examine social and obstetric factors associated with still birth and death in early infancy. The children were followed up at ages 7, 11 and 16, generally around the times of change of school institution at ages 7 and 11 and at 16 during their last compulsory year at school, theirs being the first year group for whom the minimum school leaving age was 16 years. Extensive social, education and medical data were collected at each age, a description of the 16 year data being given in Fogelman (1976). The response rate was high throughout the study, an overall response of 91, 91 and 87 per cent being obtained at each of the three ages. Goldstein, in an analysis of the characteristics of the non-respondents at 16 at previous ages in an appendix to Fogelman (1976), showed that the biases due to the complete non-respondents are small.

At present another follow up is being made of the cohort, now aged 23.

For the present study a subset of the variables at ages 7, 11 and 16 was used. At each age this included the region and characteristics of the school, information about the child's home including the numbers of brothers and sisters, social class, number of persons per room, amenities and whether the father stayed on at school after minimum school leaving age. At 7 years a multi-item description of behaviour in the home was obtained from the parents and educational information included tests of reading and arithmetic, teacher ratings of a number of attainments and information on special educational provision. At 11 years and 16 years information under each of the above headings was recorded together with, at 11 years, tests of general ability broken down into verbal and non-verbal components and of performance on a copying designs test and at 16 years an inventory of attitudes towards school.

In all 212 variables were selected. The analyses reported here concentrate mainly on the educational and home background data.

APPENDIX 3

THE CLASSICAL TEST MODEL AND THE LATENT VARIABLE MODEL COMPARED

By Russell Ecob

Two models concerned with measurement error, the Classical Test Model and the Latent Variable Model, are described, the differences highlighted and the necessary assumptions for the equivalence of the two models given.

Let x_{ijk} be a measurement of individual i on a test j at occasion k. A simple model of measurement error is that

$$x_{ijk} = \xi_{ij} + u_{ijk}$$

where u_{ijk} is the measurement error and ξ_{ij} is the 'true' value. The 'true' scores are seen to be completely defined by the measurement error giving a tautologous model (Kempf , 1980) and for the model to have any meaning it is required that the factors which are considered to contribute to the measurement error variation be explicitly defined. This we will attempt later.

The Classical Test Model (Lord and Novick, 1965, 1968) makes the following assumptions about the quantities

T1 $\underset{i \neq i'}{Cov}_{k} (u_{ij} , u_{i'j}) = 0$

T2 $\underset{k}{Cov} (u_{ij} , \xi_{ij}) = 0$

T3 $\underset{k}{E} (u_{ij}) = 0$

T4 $\underset{k}{Var} (u_{ij}) = \sigma^2_{e_j}$

T5 $\underset{k}{Var} (\xi_{ij}) = \sigma^2_j$

Note that all expectations are taken over a hypothetically infinite number of replications. Particular features of this model are that the variance of the measurement errors is assumed to be independent of the person and therefore of the true score value also. As all quantities are independent of i they hold also when summed over the population of persons. T1, called the assumption of experimental independence between items, is crucial to this and the next model and will be examined in the latter context.

*Note change in notation

The Latent Variable Model has certain differences from the classical test model. Here a vector of L latent variables, $\underset{\sim}{\eta}$, accounts for the covariation between individuals at a given point in time, K. Thus for a given item, j, and person, i, we have

$$x_{ijk} = \underset{\sim}{\lambda}_{Lj}\eta_{ij} + u_{ijk}$$

where $\underset{\sim}{\lambda}_{Lj} = (\lambda_{1j}, \lambda_{2j}, \ldots, \lambda_{Lj})$ are the loadings on the L latent variables

$\underset{\sim}{\eta}_{Li} = (\eta_{1i}, \eta_{2i}, \ldots \eta_{Li})$ and u_{ijk} is the measurement error. The foliowing definitions hold dropping the suffix, k. The latent variable values and loadings are viewed as independent of the replications, k.

L1 $\underset{i}{Cov}\,(u_{ij}, u_{ij'}) = 0$

L2 $\underset{i}{Cov}\,(u_{ij}, \eta_{Li}) = 0 \quad L = 1, \ldots, K$

L3 $\underset{i}{E}\,(u_{ij}) = 0$

L4 $\underset{i}{Var}\,(u_{ij}) = \sigma_j^2$

Here the expectations are taken over a hypothetically infinite population of individuals (i). The variances of the latent variables are fixed by fixing, say, $(\lambda_{L1}, L = 1 \ldots L)$ for person i = 1 and item j = 1.

The $\underset{\sim}{\eta}$ may comprise both general and test specific latent variables the latter having non-zero loadings only for a certain group of tests sharing a common characteristic under the unidimensional assumption, L = 1 and this will be assumed for the present discussion.

For the latent variable approach to be related to the classical test model the following condition known as the experimental independence between persons needs to be added to the latent variable model

L5 $\underset{i \neq i'}{\underset{K}{E}}\,(x_{ijk} - \eta_{ij})(x_{i'jk} - \eta_{ij}) = 0$ or $\underset{i \neq i'}{\underset{K}{Cov}}\,(u_{ijk}, u_{i'jk}) = 0$

This is the condition that the covariance of errors of any two individuals

over replications is zero, and for instance in a group test would assume no mutual influence to occur on test scores. This would not be expected to hold if cheating or other test-related mutual interaction was involved. Unless this condition holds, the commonly calculcated reliability coefficients based on a single trial based on factor or latent trait analyses will often be overestimates of the true reliability as will be shown later.

Conversely, Guttman (1945, 1969) shows that the above condition of experimental independence between persons in infinite populations of persons leads to the true scores being independent of the particular trial and thus to parallelism of trials. The assumption of experimental independence between persons is crucial in the context of assessing the dimensionality of tests and items or of the number of latent traits (McDonald, 1981).

Turning to the item domain the analogous assumption is that of experimental independence between items mentioned earlier which applies to both classical test theory (TI) and to latent trait theory (LI). This will not hold if the response to an item or test is dependent on the responses to previous items or tests. This is necessary in order that a test with a calculable reliability can be constructed by selection from a pool of items on which reliability coefficients have been independently calculated, or that internal consistency measures of reliability are consistent.

A modification to latent trait theory to allow for the relaxation of the assumption of experimental independence between items is made by Guttman (1953) and Kempf (1977). McDonald (1981) gives a full discussion of the assumption under the name local independence, in relation to the dimensionality of tests and items. Even these extension s however, do not allow the response to a given item by different persons to be differentially dependent on previous items. This could arise in test situations where a

person related variable such as fatigue or test anxiety may relate to performance in the same way for different persons but may itself be differentially affected in different people by a given item.

Reference

Kempf (1980): Paper read to session on educational applications of latent trait models at the fourth international symposium on educational testing, Antwerp, Belgium, June 1980.

APPENDIX 4

The Distribution of Variables and Transformations by Russell Ecob

A requirement of the maximum likelihood methods for estimating latent structures on multivariate data described in Chapter 2.4 is that the data are multivariate normal. When the observed data does not possess multivariate normality, transformations may, in certain cases, produce this property or an acceptable approximation. A necessary and sufficient condition for multivariate normality is that any linear combinations of the variables is normally distributed. In addition, all the normally distributed marginal variables are linearly related as are any linear combinations of these. Thus the multivariate normal distribution has strong linear properties and it is this which allows the use of reasonably straightforward linear statistical techniques. But what if no transformation of the data will produce multivariate normality? We then have a choice, given that we wish to transform the data, of either achieving linearity of relationships between variables, which can always be done by non linear transformations of individual values, or of obtaining normality of each marginal distribution separately by the same method but sacrificing the linearity of relationships between variables. We have also the further alternative of using transformations within a certain class, e.g. the one or two parameter or shifted power transformation of Box and Cox (1964). In this case neither of these desirable propoerties may hold. We therefore need to ask whether we are justified in making these non-linear transformations and if so, which property, of marginal normality or linearity should we regard as more important. First of all we distinguish non linear and parametric transformations.

Use of Non Linear Transformations and Parametric Transformations

We define non linear transformations to be those which transform particular ordinal values to a particular interval scale individually. The variable will be transformed to a particular distributional form be it one of the theoretical distributions or the distribution, maybe arbitrary,

of another variable when a linear relationship to this variable will result. We distinguish between these transformations and the parametric transformations which transform the variable through a parametric relation, for example, the transformation set of Box and Cox(1964).

The non linear transformations are equivalent to the scaling of ordered categorical data (Kendall and Stuart, 1973) and assumes that no information regarding the interval scale properties of the raw data is regarded as relevant to the analysis (indeed any interval relationships can result from such a transformation). However, the transformed data is regarded as having interval scale properties. For instance, the difference between the mth and m+r th order statistics and the n th and n + r th order statistics is assumed to have a certain value as well as a certain sign and a person who in a test of attainment taken at two occasions improves from the m + r th position to the m th position generally has a different degree of improvement from someone who improves from the n + r th position to the n th for $m \neq n$. Thus the characteristics of the distributional form to which the data are transformed are deemed to be relevant to the data, the raw scores being arbitrary apart from the ordering relation. This distinguishes the use of these transformations from the use of non-parametric techniques making no assumption on the data beyond the ordering of scores and giving an equal improvement to the two persons mentioned above. Bock(1975) takes the position that the extent of theoretical and empirical arguments for normality do not generally justify the use of non-parametric techniques apart from in small sample tests of the null hypothesis. We will generally be working with large samples and examining complex relationships and generally endorse this line.

Non Linear Transformations and their Justification

We confine the following discussion to scores on tests of attainment or ability. It is common to transform test scores to a standard normal distribution. Indeed, no test is marketed without 'standardising' on a suitable population. Therefore when using a test on a random sample from a population probably different from that on which the test is standardised, perhaps in regional and demographic characteristics and in 'up-to-date-ness', it may be justified to restandardise the observed scores to a normal distribution. In doing this we are usually saying that we regard the standardisation as inappropriate and so we cannot make any inferences on the relation of our population to the standardised one. Alternatively, we examine the relation of our population to the standardising one by comparing the standardised scores from the test manual with those from the separate standardisation of our distribution or, less strongly, compare the mean of distribution standardised according to the test standardisation with the test standardisation distribution to infer relative overall level of attainment in our population.

These transformations assume that the distribution of the appropriate attainment or ability in the population is normal.

An alternative non linear transformation is that which allots an age-specific attainment (e.g. reading age) to an individual. Here a test is given to a population of varying ages and each raw score is allotted a value corresponding to the age whose average attainment is this particular score. This will not in general give a normal distribution of scores particularly in an attainment such as reading where progress is not constant with age and where certain experiences, difficult to acquire at a particular age, may be necessary in order to achieve a certain score. Moreover this form of transformation may not be suitable when a test is given for its diagnostic as opposed to its placement value.

A further use of non linear transformations is to transform to a linear relationship with another, perhaps previously transformed variable. This may be done where further examination of the relationship between the two variables is of interest and there is no external evidence that this should not be linear.

The necessary question to ask here is why it is one variable rather than the other which is transformed. Again convenience of statistical analysis would lead in a conditional analysis to the dependent variable first being transformed to normality and then the independent variable being transformed to linearity with it. The conditional distribution of the dependent variable is then normal, though not of constant variance for all independent variable values unless this in turn is normal and the generalised least squares estimation procedure will produce optimal estimates.

A further possibility is for both variables to be transformed by canonical methods in order to maximise the correlation between them (see Kendall and Stuart, (1973), Vol 2, p 588). Here the relationship between the two transformed variables is linear though neither of the distributions have a predetermined form. However, if variables can be transformed to joint normality these will have the maximum correlation and so when this property holds, all the methods described provide the same transformation. A natural reservation about this approach is the maximisation of a quantity, the correlation which is the only quantity further analysed This describes the relationship between the variables which is often the focus of interest. However, if one is not to use this approach one has to adopt some priority on the variables included in the analysis first by fixing the distribution of one variable or alternatively by fixing the marginal distributions of all variables to known form.

A second difficulty is that of applying the canonical analysis to more than two variables. A possible interpretation is that given by Healy and Goldstein (1976) of minimising the sum of squared deviations from the assumed underlying value at a given time of several indicators, the summation being over a number of time periods. This again assumes a linear relationship between the values of the underlying variable at different times.

Linearity or Marginal Normality?

In the fields in which these non linear transformations are usually applied we generally have no theories which specify particular distributionalforms or relationships between the variables in question. We may contrast this with the case in physics where a particular law relates the height dropped from and final velocity of a mass in a vacuum. Both these quantities are measured according to measures with known interval scale properties, and the non linear transformations (say to linearity between variables) would be inappropriate (the relationship is thought to be quadratic) and the parametric transformations would only be used here in order to provide a more powerful test of the relationship given particular (ordinary least squares) statistical techniques.

Also relevant in this connection is the change of relationship from quadratic to linear when using a quadratic transformation on one of the variables. One degree of freedom seems to have been gained in the testing of the relationship and should have been taken into account in the transformation. Thus non linear transformations reduce the number of degrees of freedom in the data by the number of values on each transformed variable which are independently transformed. Thus degrees of freedom for testing only remain when a number of persons score at the same value on a test.

Transformations giving linear relationships may generally be justified if no reasons are hypothesised for the relation to be non linear or, indeed, where non linearity of relationships between variables would be uninterpretable. Thus in the case of reading attainment, when it is not anchored in terms of age equivalent scores or any other external properties, any non linearity may have no reasonable interpretation, the relation of scores between ages being completely described by the correlation. The effect of other variables, e.g. social data, home background, school characteristics on this as a criterion may then be examined.

Normality of distribution is widely encountered in naturally occuring distributions (e.g. height) and is known by the Central Limit Theorem to result

from the sum of a large number of randomly varying quantities. Thus a long test on which the response to successive items is independent should produce a normal distribution of scores. Failure to do so may be due either to the test being too short (e.g. less than 50 items), to non-independence of responses to items, or to a large proportion of the items being generally too easy or too hard. In reality all of these explanations usually hold particularly the second. However, even with items which vary in difficulty it can be argued that given a suitable choice of items a test of infinite length will have a normal distribution. A weaker argument is that the ability or attainment in question is thought to represent the sum of a large number of randomly varying influences and thus be normally distributed.

In this case the test will be regarded as having a non-normal distribution for reasons only of faulty test construction. The above argument, however, assumes a homogeneous population. What if there are, say, two sub-populations each having different mean attainments at a particular age? Under the previous argument we have a combination of two normal distributions with different means giving non-normality overall. The weaker argument of normal distribution of ability assumes that there are a very large number of such sub-populations corresponding to divisions of the overall population on different characteristics and none have differences in their means substantially larger than the others.

The most common reason for a non-normal distribution of scores in a particular test is the use on an inappropriate poulation making the test either too difficult or too easy. Then the spread of scores at one extreme of the rang is not sufficient to differentiate the assumed real difference in ability or attainment giving a a skewed distribution. This the case in the reading test at 7 years used in the NCDS and results from defective test standardisation by the test producers (as the NCDS sample is a national one and effectively random) and is to a certain extent true of the reading test at 16 years (which was standardised), originally on an 11 year sample.)

In all cases a variety of possible transformation will give linear relationships between two variables each giving differing correlations. Whereas in the bivariate normal case the appropriate transformation is obvious, in other cases we have to play off statistical convenience, which leads to the dependent variable first being transformed to normality against a desire to place limits on the degree of non normality of all the distributions concerned or to maximise the correlation. When there is more than one dependent variable it may not be possible to ensure that each is normal and at the same time allow linear relationships between them.

Why do we obtain Non-Multivariate-Normal Distributions?

We have argued that it rarely makes sense to assume non linear relationships between attainments at different occasions and that when a test has a non-normal distribution it is generally admissable to transform the scores to have a normal distribution, this then being representative of the distribution of ability or attainment in the population. Why then do we not always obtain multivariate normal distributions?

One reason has been suggested earlier. It is that the population is not homogeneous. Other explanations have to do with the nature of the ability or attainment tested and the nature of its development. It is difficult to imagine that an attainment, say, of reading, will have the same nature at different ages. At age seven the skills learnt will be more to do with the recognition of individual words, whereas later, at age 11, they will have to do more with the solving of complex syntatical problems. A word recognition test may be more appropriate to the seven year old and a reading comprehension test more appropriate to an 11 year old. However, these different attainments may require other attainments (e.g. world and subject knowledge) before they are able to form the basis for further development of reading attainment. If a skill on one attainment, say on reading, is gathered at the expense of another, say, world knowledge, then above a certain stage a high attainer at age seven may not be expected to maintain his high position relative to the rest of the sample. (Note: this is a different argument from regression to the mean) and so the relation between attainments

between occasions will be non linear. This argument can, however, also be used to justify a natural ceiling on a particular attainment at a particular age and if at both occasions natural ceilings existed, then the relationship of attainments between occasions could be again linear.

A further possibility is a defective test. This could be a test which does not correctly order the subjects on the attainment supposedly measured and which does this in a non-random way (it was seen earlier that if the error distribution was the same as the distribution of observed scores then the relationship between the true scores on two tests is the same as that between the observed scores). This could be due to particular test-related factors which affect different subjects or different subpopulations differentially. Possible examples are test anxiety where a test may cause generally high anxiety, perhaps because of unfamiliarity, and reduce disproportionally the scores in highly anxious subjects; boredom or tedium is another possibility, or the use of words which are only familiar to members of a certain subpopulation or region.

APPENDIX 5a

THEORY OF INSTRUMENTAL VARIABLES ESTIMATION

By Russell Ecob

GENERAL THEORY

Let X_{1i}, X_{2i} be the observed values of test score variables measured as deviations from their means at the first and second occasions. They are the predictor and dependent variables respectively in a simple linear regression model. Let T_{1i}, T_{2i} be their true values and e_{1i}, e_{2i} be the errors of observation of the ith subject ($i = 1, \ldots .n$).

Thus we have

$$X_{1i} = T_{1i} + e_{1i} \qquad (1)$$

$$X_{2i} = T_{2i} + e_{2i} \qquad (2)$$

and a model relating the true values at each occasion is

$$T_{2i} = \beta T_{1i} + u_i \qquad (3)$$

Let Z_i be the observed value of another variable, called the instrumental variable.

Then $$b_{IV} = \sum_{i=1}^{n} Z_i X_{2i} \left(\sum_{i=1}^{n} Z_i X_{1i}\right)^{-1} \qquad (4)$$

is called the instrumental variable estimator of the regression coefficient β (Johnston, 1972).

From (1), (2)/3) we have

$$b_{IV} = (\beta \Sigma Z_i T_{1i} + \Sigma Z_i u_i + \Sigma Z_i e_{2i}) (\Sigma Z_i X_{1i})^{-1} \qquad (5)$$

and $$b_{IV} - \beta = (\Sigma Z_i u_i + \Sigma Z_i e_{2i} - \beta \Sigma Z_i e_{1i}) (\Sigma Z_i X_{1i})^{-1} \qquad (6)$$

By letting the sample size tend to infinity, we have

$$\lim_{n\to\infty} (b_{IV} - \beta) = \lim_{n\to\infty} (\Sigma Z_i u_i)(\Sigma Z_i X_{1i})^{-1} + \lim_{n\to\infty} (\Sigma Z_i e_{2i})(\Sigma Z_i X_{1i})^{-1}$$

$$-\beta \lim (\Sigma Z_i e_{1i}) (Z_i X_{1i}) \qquad (7)$$

Given that $\lim_{n\to\infty} \Sigma Z_i X_{1i} \neq 0$, the condition for b_{IV} to be consistent is therefore

$$\lim_{n\to\infty} (\Sigma Z_i u_i + \Sigma Z_i e_{2i} - \beta \Sigma Z_i e_{1i}) = 0 \qquad (8)$$

The first term represents the covariance of the instrumental variable with the disturbances, u_i, the second the covariance with the error of observation of X_{2i}, and the third the covariance with the error of observation of X_{1i}. Attention has traditionally been focused mainly on the second two terms in this expression: indeed many reviews (for example Kendall & Stuart, 1977 Chapter 29, Madansky, 1959) limit their attention mainly to the "structural relationship" case where $T_{2i} = \beta T_{1i}$, and so the first term in (8) is identically zero. When this is the case it may often be possible to make a judicious choice of Z so that the second and third terms roughly cancel each other out.

In terms of the sample correlations, r_{Zu}, r_{Ze_2}, r_{Ze_1} between the instrumental variable Z and the disturbance and errors of measurement of X_2, X_1 respectively,

we have

$$b_{IV} - \beta = \left(\frac{\sigma_u}{\sigma_{e_1}} r_{Zu} + \frac{\sigma_{e_2}}{\sigma_{e_1}} r_{Ze_2} - \beta r_{Ze_1} \right) \frac{\sigma_{e_1}}{\sigma_{X_1}} \cdot \frac{1}{r_{ZX_1}} \qquad (9)$$

and if the reliability of X_1 is R, the expression becomes

$$b_{IV} - \beta = \left(\frac{\sigma_u}{\sigma_{e_1}} r_{Zu} + \frac{\sigma_{e_2}}{\sigma_{e_1}} r_{Ze_2} - \beta r_{Ze_1} \right) \frac{(1-R)^{\frac{1}{2}}}{r_{ZX_1}}$$

Expression (4) shows that if the predictor and dependent variables are reversed the instrumental variable estimator becomes its reciprocal.

2.2 The Efficiency of Instrumental Variable Estimators

We have

$$\mathrm{Var}(b_{IV}) = \sigma^2_{e_1} \Sigma Z_i^2 \; / \; (\Sigma Z_i X_{1i})^2$$

If b_{OLS} is the ordinary least squares regression coefficient defined as

$$b_{OLS} = \Sigma X_{1i} X_{2i} / \Sigma X_{1i}^2, \text{ then } \mathrm{Var}(b_{OLS}) = \sigma^2_{e_1} / \Sigma X_{1i}^2$$

The efficiency of b_{IV} relative to b_{OLS} is given by $\frac{\mathrm{Var}(b_{OLS})}{\mathrm{Var}(b_{IV})} = r^2_{ZX_1}$ (10)

Thus the criterion for an efficient instrumental variable is that it correlates highly with the predictor, X_1

2.3 The Use of Many Instrumental Variables

When we have p instrumental variables Z_j, j=1.....p

then $b_{IV} = (\sum_i \sum_j C_j Z_{ij} X_{2i}) / \sum_i \sum_j C_j Z_{ij} X_{1i}$. The combination of z_j which gives the most efficient estimate of b_{IV} can be found by choosing c_j so that $\mathrm{Corr}(\sum_i C_j Z_{ij}, X_{1i})$ is a maximum. The c_j are then the sample regression coefficients, b_j, of X_1 on Z_j, j=1.....p

Letting $\hat{X}_{1i} = \sum_j b_j Z_{ij}$ we obtain $b_{IV} = (\Sigma \hat{X}_{1i} X_{2i}) / \Sigma \hat{X}_{1i} X_{1i}$.

The efficiency of the instrumental variable estimator is now the square of the multiple correlation of the instrumental variable set with the predictor.

2.4 The Use of Dummy Variables as Instrumental Variables

The previous discussion has assumed the existence of instrumental variables which can be modelled as having simple linear relationships with the first occasion variable.

Two other cases can be distinguished. Firstly where an interval scaled instrumental variable has a non-linear relationship to the first occasion variable and secondly when the instrumental variable is categoric, for example measured on an ordinal or nominal scale.

In the first case the non-linear relationship can be modelled, say by a polynomial function, or the instrumental variable can be grouped into R categories. In the latter case each category can be represented in the usual way by a dummy variable. This takes the value 1 for this category and 0 for every other category.

Letting X_{1r_k}, $X_{1r_{k'}}$ be two observations on the first occasion variable which belong to the same instrumental variable category. Using the dummy instrumental variables to estimate the first occasion variable gives the estimate $\bar{X}_{1r}$ which is the mean value of all observations in category r.

Thus $$b_{IV} = \left(\sum_{r=1}^{R} p_r \bar{X}_{2r} \bar{X}_{1r}\right) \left(\sum_{r=1}^{R} p_r \bar{X}_{1r}^2\right)^{-1} \qquad (11)$$

Where $\bar{X}_{2r}$ is the mean of the X_{2r_k} in category r, and p_r is the proportion in category r. This is essentially the "Method of grouping" as introduced by Wald (1940).

Wald (1940), Neyman and Scott (1951) and Madansky (1959) have given conditions for consistency of the grouping method. Neccesary conditions are that (a) the grouping of X is independent of the errors e_1 and (b) that the denominator of the right hand side of (11) does not approach zero as the sample size tends to infinity. Random allocation to the groups, for example, would satisfy condition (a) but not condition (b). One ...

way to ensure (a) would be to know the relative ordering of the true values T_{1i}. This is difficult, however, without knowledge of the true values themselves which in general of course are unavailable.

The conditions for general instrumental variables are

$$\lim_{n\to\infty} \sum_i Z_i e_{1i} = 0 \ , \quad \lim_{n\to\infty} \sum_i Z_i X_{1i} \neq 0$$

The necessary and sufficient condition for consistency based on ordering by observed values are as follows. For any two groupings let

X_1,P_1 and $X_1(1-P_2)$ be the P_1 and $(1-P_2)$ percentiles of $f(X_1)$, the distribution of observed values. If $[u,v]$ is the shortest interval such that $P\{u < e_{1i} < v\} = 1$ i.e. if v-u is the range of e_{1i}, then b_{GP} is a consistent estimate of β if and only if $P\{X_1,P_1-v < T_1 < X_1,P_1-u\}$ $= P\{X_1(1-P_2) - v < T_1 < X_1, (1-P_2)-u\} = 0$

This means that the range of T_1 must have "gaps" at appropriate places where T_1 has a zero probability of occuring. Only if this is so can no misgrouping occur with respect to the observed X's.

It is clear that this condition cannot hold if the errors of measurement are normally distributed, due to the infinite range. However, if the range of e_{1i} is finite, careful sampling of X_1 can ensure no observations occur in the particular intervals. In particular, if the range of e_{1i} is approximately known then an approximately consistent estimate can be obtained.

The literature on grouping methods using observed values of X_1 has tended to focus on conditions for consistent estimates rather than quantifying the inconsistency of various grouping methods. The results on the NCDS data, go some way to remedying this situation for a particular data set. There has however, been work on the allocation to groups to optimise efficiency. These are summarised in Madansky (1959). These methods either assume that the variable is observed without error or use simulation procedures with very high reliabilities for X_1. An example

of the latter is Nair and Banerjee (1942) who find that a division into three groups using the extreme groups for estimation gave greater efficiency and consistency than when a two group division into equal sized groups was used.

The simulations in this case involved normally distributed errors whose standard deviations were 10% of the (constant) distance between any two adjacent true values. This gave conditions which approximated those given for consistency by Neyman and Scott. However, the reliability was 0.9999, seldom found in practice! Similar conditions are found on actual data by Madansky (1959) and contrary conclusions are found by Kendall and Stuart (1977) on simulated data by Brown (1957) of 9 normally distributed observations with normally distributed error. Here Kendall and Stuart show that the three groups method using extreme groups for estimation has higher inconsistency than the two groups method.

2.5 The Use of Instrumental Variables where there is more than one first occasion Variable

Equations (1), (3) generalise readily to p first occasion variables where the j^{th} variable $X_{1ij} = T_{1ij} + e_{1ij}$

and
$$T_{2i} = \sum_{j=1}^{P} \beta_j T_{1ij} + u_i \qquad (12)$$

We use instrumental variables Z_{jk}, $k = 1,.....n_j$ to estimate X_{1ij}. In order to obtain consistent estimates of the parameters β_j we require an analogue of the condition (8) for each predictor X_{1ij}. In order to obtain a set of efficient estimates we require the two conditions:

1. The instrumental variable set Z_{jk}, corresponding to each predictor has a high multiple correlation with the predictor.
2. The instrumental variable estimators of different predictors have low intercorrelations.

Clearly, condition 2 does not hold when the same instrumental variables are used for more than one predictor.

There is no simple analogue of the formula (10) for the efficiency for one instrumental variable, as the standard errors now depend on the variance-covariance matrix of the estimates.

APPENDIX 5b

Instrumental Variable Methods for the Estimation of Test Score Reliability

By

Russell Ecob & Harvey Goldstein
Department of Statistics & Computing
University of London Institute of Education
Bedford Way
London WC1

Introduction

In the following simple regression model,

$$y_i = \alpha + \beta x_i + u_i \qquad (1)$$

it is well known (Goldstein, 1979) that if the observed independent variable x contains errors of measurement, and if we wish to estimate the regression coefficient of the 'true' value of x, then the ordinary least squares (OLS) estimator is inconsistent. The simplest and most common model relating the true value to the observed value of x is (dropping the suffix i)

$$x = T + e \qquad (2)$$

where T is the true value, e the random error of measurement and Cov(T,e) = 0

It is supposed therefore that we wish to estimate the parameters α,β in

$$\begin{aligned} y &= \alpha + \beta T + u \qquad (3) \\ &= \alpha + \beta x + (u-\beta e) \end{aligned}$$

It is because x is correlated with $(u-\beta e)$ that the OLS estimator (b) in (1) is an inconsistent estimator of β. A consistent estimator is given by b/R, where R is known as the reliability of x and is defined as

$$R = \text{Var}(T)/\text{Var}(x) \qquad (4)$$

where Var (x) = Var (T) + Var (e)

In many situations, the value of R is very close to 1, and any adjustment to the usual estimate can be safely ignored. In other applications, for example in mental testing, R may be considerably less than 1 so that an adjustment becomes necessary. In a linear model with several further independent variables, the estimators of these too will be inconsistent if OLS is used, and consistent estimates may be obtained by adjusting the observed covariance matrix of the independent variables so that the observed variances corresponding to variables containing measurement error have estimators of their measurement error variance subtracted prior to inversion of the matrix etc., in order to calculate the coefficients (Fuller et al, 1974). To do this, it is important to have accurate and consistent estimates of the measurement error variances, or alternatively reliabilities, and in this paper we explore some new procedures for obtaining such estimates based on instrumental variable techniques.

to ensure that item responses are indeed determined by a single quantity for each individual such as given by (7). For the types of educational tests we deal with in this paper, it seems even less likely that a unidimensional trait is operating. A more detailed discussion of this topic is given by Goldstein (1980). Secondly assumption (6), often known as the "local independence" assumption, a priori seems somewhat unreasonable. It is difficult to imagine that for a given individual, if he or she fails one item then the probabilities of success on later items are the same as when he or she succeeds on the earlier item. Nevertheless, there seems to have been little, if any, serious study of this problem and the consequent effect of non-zero correlations on reliability estimates. A further discussion of this point in the context of latent trait models is given by Goldstein (1980). Thus, there is as yet no really satisfactory method for obtaining a consistent estimate of reliability using "internal" methods, nor even of providing a lower bound, and we suggest that estimates based on these methods should be treated with some caution.

1.2 External Estimates of Reliability

The most obvious method of estimating reliability or measurement error variance is to carry out repeat measurements. Thus, we have (dropping the suffix i), for two applications of a test,

$$X_1 = T + e_1 \qquad (8)$$

$$X_2 = T + e_2$$

and $$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(e_1 - e_2) = 2\left[\sigma_e^2 - \mathrm{cov}(e_1, e_2)\right]$$

For many physical measurements it is reasonable to assume independence of measurement errors, i.e. $\mathrm{Cov}(e_1, e_2) = 0$

so that we have $$\sigma_e^2 = \tfrac{1}{2}\,\mathrm{var}(X_1 - X_2) \qquad (9)$$

For mental tests, however, this usually will not be a reasonable assumption due to the presence of memory effects, learning, etc. If more than one test relating to the same thing is available, then by assuming suitable relationships between the true scores on the tests it is possible to obtain reliability estimates. The usual assumption is that the tests are congeneric so that we have, for a set of p tests,

$$X_{ji} = a_j + b_j T_i + e_{ji} \qquad j = 1 \ldots p \qquad (10)$$

The observed covariance matrix of the X_{ji} contains $\tfrac{1}{2}p(p+1)$ elements and if

we assume $\underset{j\neq j'}{Cov}(e_j, e_{j'}) = 0$ and $Cov(T_j, e_j) = \bar{0}$

the matrix is a function of the b_j and error variances $\sigma^2_{e_j}$, which gives 2p parameters. Hence, for three or more tests, unique estimates, for example maximum likelihood ones, are available. Details of this approach are given in Jöreskog (1971). Although it is not quite as serious as in the simple test-retest case, this method also has the difficulty that the measurement errors of the tests may be correlated, for example because of day to day fluctuations among examinees etc. This immediately raises the question of definition of true score, but we shall postpone a discussion of that until a later section.

In section 2 we propose a generalisation of congeneric tests to include any variable having non-zero correlation with the test whose reliability we wish to measure.

Such an 'instrumental variable' does not require any assumptions about unidimensionality or independence and also, unlike the simple test-retest or the congeneric test models, the possibility of choosing any variable means that we can search for those which are likely to be uncorrelated with the measurement error e_1. The possibility of dropping both these restrictive assumptions is attractive and the remainder of the paper investigates this problem using an extensive data longitudinal data set.

2. The Data

The data come from the National Child Development Study (NCDS) which followed up a cohort of 17 000 children born in one week of March 1958, at the ages of 7, 11 and 16. The children belonged to the first year-group for whom the minimum school leaving age was 16 years. A description of the social and educational data (among others) collected at these ages is given in Fogelman (1976).

Testing in the NCDS was carried out by the class teacher. Since the study was a national study of all children born in a particular week, most children selected were tested in a different situation and by a different tester who also scored the test.

Four possible situations giving rise to response variation are as follows:

1. The environment in which the test is administered

2. The process of test administration
3. The coding and scoring of the test (this includes the interpretation of the correctness of the response)
4. Day-to-day variation in individual test performance

Since only one test of a given type was done by each child at each occasion, the sources of variation 1-4 above are confounded. It is important, however, to distinguish 'day to day' variation from changes in true score over time.

We can regard variation over time as contributing either to measurement error or to true score variation or to both. A reasonable estimate of the true score at a particular moment would be obtained from a moving average of scores taken at successive time intervals before and after. The continuous change in true test score over, say, a week is therefore regarded as being supplemented by random error to produce the observed day to day variation. The various educational measurements in the NCDS were completed within a week for each child, so that any true score changes over not more than a one week period, are effectively regarded as part of day to day variation

In addition to these sources of measurement error, there will typically remain an unexplained variation which can be conceptualised as the variation between the response to an item and its hypothetical replication.

Cronbach et al (1972) argue that test evaluation or "generalisability" studies which also view a particular test as a sample from a universe of tests and which use experimental designs to estimate individually the above components of variation, should be carried out prior to test administration.

We use here a "Test-specific" interpretation of true score which treats true score as relevant only to the particular test. A justification for this is given by Goldstein (1979), although the methods used in this paper can be extended to a full 'generalisability' approach.

Goldstein (1979), using the same NCDS data, also drew attention to the use of instrumental variables in estimating the relation between mathematics and reading attainments, when measured at different ages. He emphasised the potential usefulness of this method when imprecise prior knowledge about the reliability of the earlier attainment scores is available, and pointed out that little was known about the degree to which the instrumental variables used satisfied the conditions of consistency.

In this paper the properties of a variety of instrumental variables are examined in the context of the regression of 16 years attainment on 11 years attainment for mathematics and reading test scores separately. Comparisons are made with

the use of ordinary least squares and also with the use of the internal estimates of the reliability coefficients for the 11 year attainment given in Goldstein (1979).

3. Theory of Instrumental Variables Estimation

3.1 General Theory

Let X_{1i}, X_{2i} be the observed values of test score variables measured as deviations from their means at the first and second occasions and let them be the predictor and dependent variables respectively in a simple linear regression model. Let T_{1i}, T_{2i} be their true values and e_{1i} e_{2i} be the errors of observation or measurement errors for the ith subject (i = 1, ...n).

Then we have, as before;

$$X_{1i} = T_{1i} + e_{1i} \qquad (11)$$

$$X_{2i} = T_{2i} + e_{2i} \qquad (12)$$

and a model relating the true values at each occasion is

$$T_{2i} = \beta T_{1i} + u_i \qquad (13)$$

Let Z_i be the observed value of another variable, called the instrumental variable.

Then
$$b_{IV} = \sum_{i=1}^{n} Z_i X_{2i} \left(\sum_{i=1}^{n} Z_i X_{1i}\right)^{-1} \qquad (14)$$

is called the instrumental variable estimator of the regression coefficient β (Johnston, 1972).

From (11), (12), (13) we have

$$b_{IV} = \left(\beta \Sigma Z_i T_{1i} + \Sigma Z_i u_i + \Sigma Z_i e_{2i}\right)\left(\Sigma Z_i X_{1i}\right)^{-1} \qquad (15)$$

and
$$b_{IV} - \beta = \left(\Sigma Z_i u_i + \Sigma Z_i e_{2i} - \beta \Sigma Z_i e_{1i}\right)\left(\Sigma Z_i X_{1i}\right)^{-1} \qquad (16)$$

In terms of the sample correlations, r_{Zu}, r_{Ze_2}, r_{Ze_1} between the instrumental variable Z and the disturbance and errors of measurement of X_2, X_1 respectively and the reliability R of X_1

$$b_{IV} - \beta = \left(\frac{\sigma_u}{\sigma_{e_1}} r_{Zu} + \frac{\sigma_{e_2}}{\sigma_{e_1}} r_{Ze_2} - \beta r_{Ze_1}\right) \frac{(1-R)^{\frac{1}{2}}}{r_{ZX_1}} \qquad (17)$$

where σ_{e_1}, σ_{e_2}, σ_u are respectively the standard deviations of the errors on the 1st occasion, 2nd occasion and the disturbance term. As the sample size tends to infinity, the following consistency condition is obtained

$$\sigma_u \rho_{zu} + \sigma_{e_2} \rho_{ze_2} - \beta \sigma_{e_1} \rho_{ze_1} = 0 \qquad (18)$$

Equation (14) shows that if the predictor and dependent variables are interchanged, the instrumental variable estimator becomes its reciprocal. We note also that the efficiency of the instrumental variable estimator with respect to the ordinary least squares estimator is $r^2_{zx_1}$ (Durbin, 1953).

3.2. The Use of Many Instrumental Variables

When we have p instrumental variables Z_j, j=1.....p

$$b_{IV} = (\Sigma_i \Sigma_j C_j Z_{ij} X_{2i}) / \Sigma_i \Sigma_j C_j Z_{ij} X_{1i} \qquad (19)$$

The combination of Z_j which gives the most efficient estimate of b_{IV} can be found by choosing c_j so that $\text{Corr}(\Sigma_i C_j Z_{ij}, X_{1i})$ is a maximum. The c_j are then the sample regression coefficients, b_j, of X_1 on Z_j, j=1.....p .

Letting $\hat{X}_{1i} = \Sigma_j b_j Z_{ij}$ we obtain $b_{IV} = (\Sigma \hat{X}_{1i} X_{2i}) / \Sigma \hat{X}_{1i} X_{1i}$.

The efficiency of the instrumental variable estimator is now the square of the multiple correlation of the instrumental variable set with the predictor.

3.3. The Use of Dummy Variables as Instrumental Variables

The previous discussion has assumed the existence of instrumental variables which can be modelled as having simple linear relationships with the first occasion variable.

Two other cases can be distinguished. Firstly where an interval scaled instrumental variable has a non-linear relationship to the first occasion variable and secondly when the instrumental variable is categoric, for example measured on an ordinal or nominal scale.

In the first case the non-linear relationship can be modelled, say by a polynomial function, or the instrumental variable can be grouped into R categories.

In the latter case each category can be represented in the usual way by a dummy variable. This takes the value 1 for this category and 0 for every other category. Let X_{1r_k}, $X_{1r_{k'}}$ be two observations on the first occasion variable which belong to the same instrumental variable category. Using the dummy instrumental variables to estimate the first occasion variable gives the estimate $\bar{x}_{1r}$ which is the mean value of all observations in category r.

Substituting in (19) gives

$$b_{1V} = (\Sigma_r p_r \bar{X}_{2r} \bar{X}_{1r}) \; (\Sigma_r p_r \bar{X}^2_{1r})^{-1} \qquad (20)$$

Where $\bar{X}_{2r}$ is the mean of the X_{2r_k} in the category r, and p_r is the proportion in category r. This is essentially the "Method of grouping" as introduced by Wald (1940).

The literature on grouping methods (for example Wald (1940), Neyman and Scott (1951), Madansky (1959)) using observed values of X_1 has tended to focus on conditions for consistent estimates and on the relative efficiency of different groupings rather than quantifying the inconsistency of various grouping methods. The results on the NCDS data given below, go some way to remedying this situation for a particular data set.

4. Application of Instrumental Variable Methods to the NCDS Data

4.1 Selection of Variables

In all, 50 variables are considered as instrumental variables, being measured at ages 7, 11 and 16. These consist of test scores, teacher ratings and background variables. The test scores are of reading and mathematics at each age and in addition of general ability and copying designs scores at age 11. The teacher ratings are of reading and mathematics at all ages, and in addition of oral ability and creativity at age 7, of oral ability, and general knowledge at age 11 and of practical subjects at age 16. The "background" variables are social class and indices of behaviour in the home at all three ages; the number of children in the household, birth order of the child, an index of accommodation facilities and childrens' heights at ages 11 and 16, overcrowding at 11; and region, indices of school behaviour, and a variety of feelings towards school at 16. The reader is refereed to Davie, Butler and Goldstein (1972), and Fogelman (1976) for a more complete description of these variables.

The variables used as predictor and dependent variable in the regressions, that is the test scores at age 11 and 16 of mathematics and reading, are transformed to have standard normal distributions in the same was as in Goldstein (1979) who showed that near linear relationships between observed scores resulted.

As the relative efficiency of an instrumental variable estimator is proportional to the correlation with the predictor, variables are only retained

for further analyses when this correlation is greater than 0.3. This eliminates most of the "background" variables but only one of the teacher ratings' (that of outstanding ability in any area at age 11) and none of the test scores, leaving 25 variables in all. All cases with missing values on any of these 25 variables are excluded leaving 5371 cases with test scores at each age.

In the appendix to Vogelman (1976), Goldstein shows that the attrition of subjects in the study does not affect to any marked extent the relationships found and Goldstein (1979) finds that test scores for subjects having missing values on the background variables show no significant differences from other subjects.

In the results reported here, all instrumental variables are treated as sets of dummy variables for reasons given in Section 2.4. For the test scores the dummy variable coding into five roughly equal size groups ensures that the relative loss in predictive efficiency from dummy variable compared to simple linear regression is always less than 6%.

Using these instrumental variables as dummy variables or as interval scale variables in fact gives very simlar results, the maximum difference in regression coefficients for any instrumental variable being 2%.

4.2 Forming hypotheses of error structure in prediction relations

As in Section 1.2 we assume that the true score of a test comprises that component of test score which is unaffected by day to day variation by the particular tester or by the test situation, and is specific to the particular test used. We now examine the correlation between the variables to be considered as instrumental variables with measurement errors on the first occasion and second occasion tests and with the disturbance term. These variables are teacher ratings on a variety of attainments, test scores and social class.

The teacher ratings, like the test scores, in general will contain measurement error, thus reflecting and being reflected by variations in the child's interest in subjects and day to day variations in the type of relationship to the teacher. Thus a teacher who has very recently seen a child do a good piece of work or show a keen interest, may tend to rate him higher than otherwise. If the same contributory factors affect test score then a teacher rating made at the same time as the test would be expected to have a positive correlation with the test score measurement error. Likewise where a different attainment is rated at the same time as the test similar correlations may exist, although presumably smaller.

There are two variables which we hypothesies will have a zero correlation with test score measurement error. These are teachers ratings taken at a different point in time and social class. We would expect none of the sources of measurment error to relate to teacher rating at a point in time 4 or 5 years away. Nor would we expect social class, which does not vary very much for an individual over short time periods, to relate to any of the sources of measurement error.

Finally we examine the relation of the disturbances u_i, to teacher ratings and social class. Either of these variables and particularly social class may be correlated with the disturbances if they relate to the dependent variable once the predictor variable has been controlled for.

The hypotheses formulated above may be summarised thus;

H1. Teacher ratings on a test where the rating is at the same time as the test, will be positively correlated with test score measurement error.

H2. Teacher ratings on a different attainment from that tested, where the rating is at the same time as the test will be postively correlated with test score error but to a lesser extent than for the same attainment.

H3. Teacher ratings when the child is at a different age from that of the test will be uncorrelated with test score measurement error whether or not the same attainment is tested.

H4. Teacher ratings are not correlated with disturbance terms from the regression of second occasion score on first occasion score.

H5. Social class is correlated with disturbance terms.

H6. Social class is not correlated with test score measurement error.

Hypotheses H1, H2, H3 refer to the relation of teacher ratings to test score measurement error. H1, H2 refer to teacher ratings at the same time as the test and H3 at a different time. H4 refers to the relation of teacher ratings to equation disturbances. H5, H6 refer to social class and apply to the relation with equation disturbances and test measurement errors respectively. Generally, teacher ratings are held to correlate with test score measurement errors only when tested at the same time as the test and not to be correlated with equation disturbances. In contrast, social class is hypothesised as correlating with disturbances but not with test score measurement error even when measured at the same time as the test.

Examining (17) these six hypotheses give rise to the following predictions. The hypotheses giving rise to each prediction are given in brackets after

the prediction.

P1. Comparing teacher ratings at 11 years, the lowest estimate of β will occur for the teacher rating of the same attainment, (from H_1 to H_4, particularly H_2 affecting r_{Ze_1}), and at 16 years the highest value will occur for the rating of the same attainment (from H_1 to H_4, particularly H_2 affecting r_{Ze_2}).

P2. For a given attainment, teacher ratings will give higher estimates of β when measured at 16 than 11 years (from H_1 to H_4).

P3. For a given attainment, teacher ratings will give higher estimates of β when measured at 7 rather than 11 years (from H_1 to H_4).

P4. There is no difference in estimates between teachers ratings at 7 years (from H_3, H_4).

P5. Social class gives higher estimates of β than teacher ratings at 7 and 11 years (from H_1 to H_6).

P6. Social class will give similar estimates of β irrespective of the age at which measured (from H_5, H_6).

It should be noted that these predictions are not unequivocal tests of the hypotheses. For instance, even if P6 holds one could conceive of different correlations of social class with measurement error at different ages, these terms being counteracted by different correlations with equation disturbances. This, however, seems unlikely.

4.3 Results

Tables 1 and 2 give estimated regression coefficients for reading and mathematics respectively for 16 year attainment on 11 year attainment using a variety of teacher ratings and social class as instrumental variables at ages 7, 11 and 16. Using the ungrouped 11 year test score as instrumental variable gives the ordinary least squares estimate, which thus enables reliability estimates to be calculated for each choice of instrumental variable by dividing the ordinary least squares estimate by the instrumental variable estimate. Each prediction will be examined in turn.

Table 1 Estimated Regression Coefficients of Reading Test at 16 years on Reading Test at 11 years adjusted for Measurement Error, Using a Number of Instrumental Variables Separately

Instrumental variable measured at:	7 years	11 years		16 years	
Teacher rating of:					
Oral	0.955	Oral	0.972		
Reading	0.944	Use of Books	0.964	English	1.042
Number	0.990	Number	0.974	Mathematics	1.067
Creativity	0.990				
		General Knowledge	0.979	Practical Subjects	1.047
Social Class	1.057	Social Class	1.070	Social Class	1.064
Reading Test at 11 (interval scale)(OLS estimate)	0.797				
Reading Test at 11 (5 category groupings)	0.810				

Average standard errors of regression coefficients:	Using Teacher ratings at 7 years	0.017
	" " " " 11 "	0.013
	" " " " 16 "	0.015
	" Social Class	0.027
	" Reading Test at 11 years	0.008

Table 2 Estimated Regression Coefficients of Mathematics Test at age 16 on Mathematics test at age 11 adjusted for measurement error using a number of instrumental variables separately

Instrumental variable at:	7 years		11 years		16 years	
Teacher rating of:						
Oral	0.883	Oral	0.890			
Reading	0.849	Use of Books	0.884	English	0.992	
Number	0.854	Number	0.874	Mathematics	1.073	
Creativity	0.911					
		General Knowledge	0.920	Practical Subjects	1.029	
Social Class	0.994	Social Class	0.991	Social Class	1.025	
Maths Test at 11 (Interval Scale) (OLS estimate)	0.748					
Maths Test at 11 (5 category grouping)	0.763					

Average standard errors of regression coefficients:

Using Teachers ratings at 7 years	0.018
" " " " 11 years	0.015
" " " " 16 "	0.016
" Social Class	0.030
" Mathematics Test at 11 years	0.010

P1. This holds for mathematics using teacher ratings both at 11 and 16 and for reading for teacher ratings at 11 but not 16.

P2. This holds for comparable teacher ratings at 11 and 16 for both reading and mathematics test scores.

P3. This only holds for one out of the six possible comparisons, namely teacher rating of "number" for reading attainment regression.

P4. This holds for both attainments.

P5. This holds in all cases.

P6. This holds at all ages for both attainments.

It should be noted that predictions P4 and P6 specify no differences between regression coefficients whereas the other predictions are of a difference in a specified direction. In fact, for P4, P6 the differences between coefficients are small in relation to the standard errors.

The predictions are all seen to hold generally with the exception of P3. This implies the rejection of H3 or H4 or both. Rejecting H3 implies that the differences in coefficients among the different teacher ratings at age 7 should be similar to those using the three corresponding teacher ratings at age 11. This is true with one exception, this being the reversal in the relative magnitude of the teacher ratings of reading and number between ages 7 and 11 for mathematics.

The smaller coefficient estimates using 7 year rather than 11 year ratings could be explained by a correlation of 11 year rating with errors in the dependent variable, which counteracts the correlation with errors in the independent variable - the 7 year ratings having lower correlations with the dependent variable.

If H4 is the sole reason for the failure of P3 this suggests that the partial correlation of teacher ratings with social class at 11 given test score at 11, would be higher for 7 year teacher ratings than for 11 year teacher ratings, and this is not the case.

It seems then that we should discard social class as a suitable instrumental variable due to its correlation with the equation disturbances, and teacher ratings at 16 years are positively correlated with test score error at 16 (from H1, H2) and probably also with equation disturbances. This leaves a choice between teacher ratings at 7 and 11 years. As H3 does not hold we cannot be completely content with using the same-attainment 7 year teacher ratings, and in addition it is not known how highly correlated these are with the disturbances. It was suggested earlier, however, that we should expect the 11 year ratings to be more highly correlated with the disturbances.

As 112 also holds, the wisest choice would seem to be the rating of a different attainment (out of mathematics or reading) at age 7. For the reading attainment this gives a reliability of 0.81 (using teacher rating of number at age 7) and for mathematics attainment gives a reliability of 0.89 using teacher rating of reading at age 7. In fact the choice between 7 and 11 years for the instrumental variable makes a little difference for mathematics attainment giving a reliability of 0.86 using reading rating at age 11, and for reading attainment the difference in reliability estimate is only 0.005.

The question naturally arises here as to whether the use of tests of the same attainment at different ages is necessary in order to obtain reasonable estimates of reliability coefficients by this method. Estimates of the reliability coefficient of 11 year reading test score obtained by regressing 16 year mathematics score on 11 year reading score using separately as instrumental variables teachers ratings of reading and mathematics attainments at 11 years, gave reliability estimates of 0.76 and 0.61 respectively compared with the value of 0.81 given above. This suggests that the disturbance terms in this regression are correlated with the mathematics teachers ratings. For the regression of 16 year reading test on 11 year mathematics test using teachers ratings of mathematics and reading separately as instrumental variables reliability estimates of 0.85, 0.68 respectively are obtained, compared with the value of 0.88 given above. Care should therefore be taken when estimating reliabilities by this method to use similar attainments as dependent and predictor variables.

4.4 Use of grouped first occasion variable as instrumental variable

Table 3 gives estimated regression coefficients for both reading and mathematics when the first occasion variable is grouped into 2, 3, 5 or 7 equal groups.

The inconsistency of the grouping estimator (b_G) relative to the ungrouped (OLS) estimator (b_{OLS}) is given by

$$k = \frac{b_{IV} - b_G}{b_{IV} - b_{OLS}} \qquad (21)$$

where b_{IV} is the instrumental variable estimator (using an appropriate teacher rating) and is assumed to be consistent.

As suggested in 3.3 substantial inconsistencies are indicated in these data, being greater with a larger number of groups. Thus there is a trade-off between inconsistency and efficiency, the latter being greater as the number of groups is increased. For reading attainment the lowest estimate of reliability, arising from the division into two groups, is 0.97. This is higher than any of the values derived from the regression estimates by instrumental variables methods given in Table 1. Furthermore, the estimated regression coefficient is seen to vary with the point of dichotomy. Whereas for reading the lowest estimates occurs for both extreme divisions, for mathematics the lowest estimate occur when division is at the lower end of the scale and the highest estimate when division is at the higher end.

Table 3 Estimated regression coefficients and standard errors using the grouped predictor as instrumental variable. Standard errors in brackets. k is defined in (21).

	Reading	k	Mathematics	k
Ungrouped (O.L.S.)	0.797 (0.0082)	1.00	0.748 (0.0097)	1.00
No. of equal groups (of equal size)				
7	0.808 (0.0085)	0.94	0.755 (0.0100)	0.93
5	0.810 (0.0086)	0.93	0.763 (0.0102)	0.86
3	0.818 (0.0092)	0.89	0.777 (0.0109)	0.73
2	0.827 (0.0103	0.84	0.780 (0.0121)	0.70
Varying position of dichotomy				
Proportion in lower test group				
0.2	0.810 (0.0012)	0.93	0.687 (0.014)	1.58
0.4	0.823 (0.0010)	0.87	0.765 (0.012)	0.84
0.6	0.822 (0.0010)	0.87	0.802 (0.012)	0.49
0.8	0.789 (0.0012)	1.04	0.805 (0.014)	0.46

If the assumption is made that the correlations of the grouped first occasion variable with the disturbances and error in the second occasion variable are both zero then using the reliability estimates from the previous section we can substitute in (26) to obtain the correlations with the error in the first occasion variable, r_{Ze_1}. These are given below in Table 4, and assume the reliability estimates given in Section 3.3 (0.81 and 0.89) are correct.

Table 4 Correlations of dichotomised instrumental variable and first occasion measurement error for different division points

	Proportion below division point				
	0.2	0.4	0.5	0.6	0.8
Reading	0.280	0.302	0.329	0.305	0.324
Mathematics	0.390	0.226	0.233	0.128	0.120

Thus the correlations, while reasonably constant for reading are systematically decreasing for mathematics. We have no good explanation for this but possible causes are non-homogeneity of errors in the mathematics test, or a non zero correlation between true score and errors of measurement.

4.5 The use of test score as an instrumental variable

Test scores of reading and mathematics at 7, 11 and 16 years, a score of General Ability at 11 years (with Verbal and Non-verbal components) and a Copying Design Test, are considered as instrumental variables, and Results are given in Table 5.

Table 5 Regression Estimates of 16 year attainment test on 11 year attainment test in reading and mathematics using test scores as instrumental variables. Standard errors in brackets

	Reading	Mathematics
Instrumental variable		
at 7 years:		
Reading Test	0.920 (0.015)	0.822 (0.017)
Mathematics Test	1.004 (0.020)	0.850 (0.018)
at 11 years		
Reading Test	0.810 (0.009)	0.866 (0.014)
Mathematics Test	0.935 (0.012)	0.763 (0.010)
General Ability Test: Verbal	0.942 (0.012)	0.826 (0.013)
: Non-Verbal	0.982 (0.014)	0.901 (0.014)
: Overall	0.957 (0.012)	0.860 (0.013)
Copying Designs Test	0.989 (0.032)	0.933 (0.032)
at 16 years		
Reading Test	1.197 (0.013)	0.949 (0.015)
Mathematics Test	1.053 (0.015)	1.315 (0.017)

Since short term fluctuations in attainment to some extent will be correlated over all attainments we would expect the arguments and predictions in Section 2.6 in relation to teacher ratings to apply to test scores. In all applications considered here the test score is divided into 5 groups, and dummy variables used. P1 is satisfied trivially in the light of the results on the use of a grouped predictor as instrumental variable. P2 is satisfied, but P3 is again contradicted by the behaviour of the reading tests at 7 and 11 when used as instrumental variable for mathematics attainment.

Generally, the behaviour of test scores is similar to the teacher ratings and the standard errors are similar, giving little indication for preference of one set of variables over the other. Nevertheless, if correlations between instrumental variables when measured at 7 years, and errors of prediction and of measurement at 11 and 16 years were zero, then similar estimates of regression coefficients should be found for the tests and teacher ratings of different attainments, used as instrumental variables. In fact, for both tests and teacher ratings and for both attainments the estimates are larger for mathematics than for reading, when used as instrumental variables.

The results for general ability indicate that it occupies an intermediate position between the two attainments.

4.6 Separate equations for each social class

Table 6 gives the estimated regression coefficient of 16 year reading test score on 11 year reading test score separately for each social class measured at 11 years, using 11 year teacher rating of number as instrumental variable in each case.

Table 6 Estimated regression coefficients for reading test at 16 on reading test at 11 for each 11 year social class separately using teacher rating of number at 11 as instrumental variable

Social Class	Regression Coefficient	Standard Error	Raw Corr. between 11 & 16 yr. scores	Estimated true corr. between 11 & 16 yr scores	Mean reading score of 11 years	Mean reading score of 16 years	No. Cases	11 year test score		16 year test score	
								Reliability	Variance of measurement error	Reliability	Variance of measurement error
Professional	0.918	0.084	0.72	0.98	0.741	0.688	249	0.75	0.193	0.72	0.199
Intermediate	0.893	0.038	0.77	0.95	0.485	0.479	903	0.84	0.138	0.78	0.175
Skilled non-manual	0.902	0.046	0.78	0.97	0.360	0.384	497	0.80	0.177	0.81	0.142
Skilled manual	0.980	0.023	0.80	0.97	-0.099	-0.099	2235	0.79	0.165	0.85	0.116
Semi-skilled manual	0.981	0.037	0.78	0.95	-0.238	-0.281	880	0.78	0.169	0.84	0.132
Unskilled manual	1.037	0.074	0.78	0.90	-0.567	-0.458	282	0.76	0.198	0.94	0.132
All (excluding no male head)	0.974	0.014	0.80	0.96	0.046	0.035	5046	0.82	0.168	0.85	0.135

Test* for equality of 11 year Reliability Coefficients $\chi^2_5 = 11.9$ $(p < 0.05)$
Test for equality of 11 year Measurement error variances $\chi^2_5 = 9.5$ $(p > 0.05)$
Test for equality of true correlation coefficients $\chi^2_5 = 141.7$ $(p < 0.001)$

*Test values for these tests are obtained from the robust chi-square test in Layard (1973) assuming the distribution of the measurement errors has a kurtosis of 3. When a kurtosis of zero is assumed the values of χ^3_5 for equality of reliability coefficients and measurement errors respectively are 29.8, 23.7 $(p < 0.001)$

With the exception of the "professional" class the regression coefficients systematically decrease with higher social class. The reliability estimates of the reading test at 11 years in each social class are seen to vary, generally increasing with higher social class. The exception to this is in the "professional" class whose lower reliability may be explained by the presence of higher measurement errors at the top of the reading scale. The measurement error variances vary in an inverse fashion and this table does not provide evidence that the measurement error variance is more nearly constant between social classes than the reliability coefficient, as suggested in Goldstein (1979), indeed the opposite seems to be the case. The estimated values of the true correlation coefficient ($\hat{\rho}$) within each social class is obtained from the equation

$$\hat{\rho} = \frac{r}{\sqrt{\hat{R}_{11}\,\hat{R}_{16}}}$$

where r is the observed correlation coefficient and $\hat{R}_{11}$, $\hat{R}_{16}$ are the estimated reliability estimates of the 11 and 16 year test scores, the 16 year estimates being obtained by using instrumental variable estimates of the regression of 11 year on 16 year test scores, the intrumental variable being the teacher's rating at age 16 of mathematics. The estimated correlations vary with social class with the "professional" class having the highest value and the "unskilled manual" class the lowest value. These values seem rather high, but it must be remembered that the same reading test was used at 11 and 16 years. A 95% confidence interval level for the overall value is (0.962, 0.996).

Table 7 gives the coefficients for the mathematics test, where teacher's ratings of the use of books at 11 years is used as the intrumental variable for each class. Contrary to the reading scores, the regression coefficients increase with higher social class, . . The reliability estimates of the mathematics test at 11 years are seen to vary with a lower value for the "professional" class. The measurement error variances vary in an inverse fashion. The extent of variation between social classes of the reliability and measurement error variances is larger than for the reading test, both reliability and measurement error variance estimates showing a similar degree of variation.

The estimated true correlation coefficients show the same pattern as for the reading test, with the values on the "professional" and "unskilled manual" classes being respectively higher and lower than the average value. The 95% confidence interval for the overall value is (0.899, 0.908).

Table 7 Estimated regression coefficients for mathematics test at 16 on mathematics test at 11 for each 11 year social class separately using teacher rating of use of books at 11 as instrumental variable

Social Class	Regression Coefficient	Standard Error	Raw Corr. between 11 & 16 yr. scores	Estimated true corr. between 11 & 16 yr scores	Mean Maths score at 11 years	Mean Maths score at 16 years	No. Cases	11 year test score		16 year test score	
								Reliability	Variance of measurement error	Reliability	Variance of measurement error
Professional	1.081	0.120	0.73	0.94	0.778	0.716	249	0.72	0.210	0.83	0.146
Intermediate	0.920	0.042	0.73	0.88	0.454	0.510	903	0.83	0.136	0.84	0.138
Skilled non-manual	0.883	0.052	0.73	0.88	0.330	0.388	497	0.84	0.132	0.82	0.153
Skilled manual	0.869	0.026	0.71	0.89	-0.063	-0.098	2235	0.85	0.118	0.74	0.215
Semi-skilled manual	0.824	0.044	0.65	0.88	-0.248	-0.226	880	0.79	0.160	0.70	0.230
Unskilled manual	0.689	0.068	0.64	0.83	-0.478	-0.475	282	0.87	0.101	0.67	0.229
All (excluding no male head)	0.885	0.016	0.74	0.90	0.039	0.071	5046	0.85	0.133	0.79	0.194

Test* for equality of 11 year reliability coefficients $\chi^2_5 = 34.5$ $(p < 0.001)$

Test* for equality of 11 year measurement error variances $\chi^2_5 = 29.5$ $(p < 0.001)$

Test for equality of true correlation coefficients $\chi^2_5 = 46.2$ $(p < 0.001)$

*See note on table 6

5. Summary and Discussion

Our results suggest that differing correlations of instrumental variables with measurement errors account for the observed differences in regression and reliability estimates, although social class has a negligible correlation with measurement error but a non-negligible correlation with the error of prediction.

The estimated correlation coefficient between true scores on reading and mathematics tests at eleven and sixteen years is respectively 0.96 and 0.90. The estimated reliabilities using the selected instrumental variables (teacher ratings of an unrelated ability at age 7 as the predictor) are 0.81 and 0.88 for reading and mathematics respectively. These compare with the values of 0.82 and 0.94 given in Goldstein (1979) by split half item analysis on a subsample of 300 cases. Whilst the values for reading are similar the value obtained by item analysis for mathematics is somewhat higher than any obtained for the instrumental variables used here, although the difference is of the same order as the standard error of the separate estimates.

For reading attainment the estimated standard error of the reliability estimate obtained by item analysis is 0.030 while the instrumental variable method gives 0.012. A split half estimate using all available data would have a standard error of about 0.007. For mathematics the relevant standard errors are 0.020, 0.014 and 0.005.

Using a grouping of the predictor variable itself as an instrumental variable gives estimates of the reliability which are higher than any obtained using other variables as instrumental variables, irrespective of the number of groups used. These estimators, we suggest, are not to be recommended.

The examination of reading and mathematics attainment separately in each social class gave different reliabilities for the different social classes.

For both attainments the "professional" social class gave the lowest value but otherwise for the mathematics test lower reliabilities occurred in the lower social classes and vice versa for the reading test. Measurement error variance estimates varied in an inverse fashion. Instrumental variable estimates of regression coefficients showed an opposite trend for each attainment, with the coefficients for the reading test being generally higher in the lower social classes and for the mathematics being higher in the higher social classes. Estimates of the true correlation between 11 and 16 years was different in each social class for both reading and mathematics attainments, with higher values in the "professional" class and lower values in the "unskilled manual" class for each attainment.

While instrumental variable estimation has had a long history (early papers on theory and application in the economic field includes Wald (1940), Reiersol (1945), Durbin (1953), Sargan (1958), Madansky (1959)), it has not yet become generally accepted as an estimation method in the social and educational fields. Sargan (1958, page 396), in discussing the (unknown) correlation of instrumental variables with measurement error in an economic context states "It is not easy to justify the basic assumption concerning these errors, namely that they are independent of the instrumental variables. It seems likely that they will vary with a trend and with the trade cycle. In so far as this is true the method discussed here will lead to inconsistent estimates of the coefficients. Nothing can be done about this since presumably if anything were known about this type of error, better estimates of the variables could be produced. It must be hoped that the estimates of the variables are sufficiently accurate, so that systematic errors of this kind are small." We have argued that comparisons of different instrumental variables, considered separately, can throw some light on the error structure in the data and thus lead to better knowledge of the consistency of the estimates produced. Furthermore it is also our view that this approach provides a flexible tool for an empirical study of the various assumptions needed to produce good estimates.

Finally, four issues seem particularly worthy of attention:

1. Obtaining estimates of the standard errors of the difference between different instrumental variable estimates (these will be lower than those obtained using the individual standard errors and assuming independence). This would enable a more careful analysis of the hypotheses of the paper.
2. Obtaining good estimates of the standard errors of the reliability estimates produced by instrumental variable methods.
3. Examination of the use of more than one instrumental variable in connection with a single predictor in terms of the efficiency and consistency of estimates.
4. The study of differing reliabilities and measurement error variances in different groups of variables such as social class, in order to incorporate these into linear model estimates.

REFERENCES

ANDERSON, E B (1980) Comparing Latent Distributions. Psychometrika, 45, 121-134

BROWN, R L (1957) Bivariate Structural Relation, Biometrika, 44, 84-90

CRONBACH, L J, GLESER, G C, NANDA, H & RAJARTNAN (1972) The dependability of behavioural measurements: theory of generalisability for scores and profiles. Wiley, New York

DAVIE, R, BUTLER, N R & GOLDSTEIN, H (1972) From Birth to Seven: a report of the National Child Development Study. Longmans, London

DURBIN, J (1953) Errors in Variables. Review of Institute of International Statistics Vol 22, 1954, 23-54

FOGELMAN, K (1976) Britain's Sixteen Year Olds. National Children's Bureau, London

GOLDSTEIN, H (1979) Some Models for Analysing Longitudinal Data on Educational Attainment. J.Roy.Statist.Soc.,A,142, 402-442

GOLDSTEIN, H (1980) Dimensionality, Bias, Independence, and Measurement scale problems in latent trait test score models. Br.J.Math. & Statist. Psychol., 33, 234-246

JOHNSTON, J (1972) Econometric Methods. McGraw Hill, New York

JORESKOG, K G (1971) Statistical Analysis of sets of congeneric tests. Psychometrika, 36, 109-133

KENDALL, M G & STUART, A (1977) The Advanced Theory of Statistics. Volume 2, Griffin, London

LAYARD, M W J (1973) Robust Large Sample Tests for Homogeneity of variances. J.Amer.Statist.Assn. 68, 195-198

LORD, F M & NOVICK, M R (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, Mass.

McDONALD, R (1981) The Dimensionality of Tests and Items. Brit.J.Math and Statist.Psychol. 34, 100-117

MADANSKY, A (1959) The fitting of straight lines when both variables are subject to Error. JASA, 54, 173-205

NAIR, K R & BANERJEE (1942) A note of fitting straight lines if both variables are subject to error. Sankya, 6, 331-333

NEYMAN, J & SCOTT, E L	(1951)	On certain methods of estimating the linear Structural Relation. Ann.Math.Statist. 22, 352-355
REIERSOL, O	(1945)	Confluence Analysis of Means of Instrumental sets of Variables. Arkiv. for Mathematik, Astronomi Och Fysik, Vol 32
SARGAN, J D	(1958)	The estimation of Economic Relationships using Instrumental Variables. Econometrica, 26, 393-415
WALD, A	(1940)	Fitting of straight lines if both variables are subject to error. Ann.Math.Statist. Vol 11, 284-300

Summary

The method of Instrumental Variables is suggested as an alternative to traditional methods for estimating the reliability of mental test scores, and avoids certain drawbacks of these methods. The consistency and efficiency of the instrumental variable method are examined empirically using data from the British National Child Development Study in an analysis of 16 year and 11-year -old scores on tests of mathematics and reading.

Keywords

Instrumental Variables, Errors in Variables, Reliability, Longitudinal, Educational Attainment.

Acknowledgements

We are grateful to the National Children's Bureau for permission to use the data and to Dougal Hutchinson for his comments on an early draft of the paper. The work was carried out largely on a grant from the National Institute of Education, Washington (NIE-G-77-0065).

APPENDIX 6 SOME APPLICATIONS OF LISREL TO EDUCATIONAL DATA

BY RUSSELL ECOB

Paper read to the Seminar on Structural Equation Modelling with Particular Reference to LISREL at the University of London Institute of Education 14th September 1981.

NOT FOR REPRODUCTION WITHOUT PERMISSION

SUMMARY

Examples are given of the use of LISREL in both confirmatory and exploratory modes. In the confirmatory mode a sequence of hypotheses is set up each of which is a special case of the preceding one and is tested sequentially. In the exploratory mode the data in conjunction with knowledge of the subject matter is used at any stage to determine which parameter to add to a given model. This paper emphasises, when used in a confirmatory mode, the importance of the initial specification of the model and shows how more than one model can be "confirmed" by the data. In the exploratory mode LISREL is used on the data from the National Child Development Study to estimate the correlation between the latent variables corresponding to underlying reading attainments at two ages. Examined also is the effect of different fitted error correlations on the correlation between latent variables and the effect on final estimates of choosing different indicator variables and of scaling variables to have normal distributions. The use of data on reading attainment at three ages allows the incorporation of extra latent variables corresponding to test specific factors. Finally an alternative method using instrumental variables for estimating change in reading attainment is briefly described and compared with the linear structural relations method.

USE IN A CONFIRMATORY MODE

Joreskog and Sorbom (1977) describe the procedure thus:

"Suppose H_o represents one model under given specifications of fixed, free and constrained parameters. To test the model H_o against any more general model H_1, estimate them separately and compare their χ^2. The difference in χ^2 is asymptotically a χ^2 with degrees of freedom equal to the corresponding difference in degrees of freedom. In many situations, it is possible to set up a sequence of hypotheses such that each one is a special case of the preceding and to test these hypotheses sequentially".

Werts, Breland, Grandy and Rock (1980) apply LISREL to the measurement of writing ability at three occasions, by essay and by a test of standard written English at each occasion, the measures being made on American undergraduates

during their first year of study, composition instruction being given between the first two occasions. They aim to use LISREL to separate out the instability or variation in the "true scores" (underlying or latent ability) over occasions from the measurement errors which are supposed to be correlated across occasions and thus derive estimates of reliability. It is this correlation between measurement errors across occasions which makes the model more general than the well known factor analysis model.

Each essay is marked independently on a scale from 1 to 6 by two outside teachers to give a scale from 1 to 12 and the tests each consist of 50 multiple choice items. The data consist of 234 out of a total of an initial 2500 students tested at each occasion and though it gives rise to a large potential non-response problem is used here for didactic purposes. The correlation matrix is given in Table 1.

TABLE 1

Werts et. al. data on writing ability of undergraduates.

Test 1	1.000					
Test 2	0.837	1.000				
Test 3	0.854	0.842	1.000			
Essay 1	0.621	0.640	0.602	1.000		
Essay 2	0.602	0.636	0.551	0.564	1.000	
Essay 3	0.596	0.617	0.597	0.572	0.523	1.000
	Test 1	Test 2	Test 3	Essay 1	Essay 2	Essay 3

Their initial (most general) model is that the true scores of the tests and essays are linearly related, the same latent variable being measured by both indicators, and that the errors of measurement of the tests are uncorrelated across occasions whereas the errors of measurement of the essays are correlated across occasions. We call this the "essay error correlation model".

An alternative hypothesis, not examined by Werts et al, is that the earrors on the tests are correlated between occasions but that the essary errors are not correlated between occasions. We call this the "test error correlation" model. The further alternative, that the true scores underlying the tests and essays represent different latent variables and that neither of the errors are correlated across occasions, leads to an unidentifiable or indeterminate model having too many parameters to be fitted by the 21 independent observed correlations. Note, however, that Block & Saris (1981) provides a reinterpretation of the Werts et al data using two latent variables having correlation 0.89 by making the assumptions that a simplex relation holds between each set of latent variables and that at least one of three other possible restrictions hold. Appendix 2 shows that the model with both sets of errors correlated across occasions is also not identified.

The choice between the "essay error correlation" and the "test error correlation" models cannot be made empirically as neither is a specialisation of the other, the two models embodying different conceptualisations of the true score or latent variable or equivalently of the allowed components of error. They will be shown to lead to different parameter estimates, in particular, reliability estimates, by carrying through the process of Werts et al for both models.

The most general model considered by Werts et al has the restriction on true scores that the correlation between the true scores at occasions 2, 3, is unaffected by the true score at occasion 1, or $\rho_{23/1} = 0$ This is equivalent to the restriction that $\rho_{13} = \rho_{12}\rho_{23}$ and is called the simplex model of true scores. The model is shown diagrammatically in figure 1 for the "test error correlation" case. The LISREL specification of this model is given in Joreskog and Sorbom (1977), Section 4.2 and also in Appendix 1 of this paper.

The model with unrestricted error correlations can be reformulated as one with a test specific factor (see Appendix 3) and this is shown in figure 2. In testing the series of possible models the restrictions on the relationship between the true scores can be viewed as the structural component of the model and we have in order of increasing restriction,

S1 No restriction on true scores

S2 Simplex restriction on true scores

S3 No change in true scores between occasions i, j

S4 No change in true scores overall

Similarly considering the measurement component given by the error correlations (for a particular test) we have the following possibilities (restrictions apply to both tests) :

E1 No restriction on error variances or covariances

E2 Equal error variances on two occasions

E3 Equal error variances on all occasions

E4 Equal error correlations between two sets of occasion

E5 Equal error correlations between all occasions

E6 Zero error correlations between all occasions

We proceed by testing first the structural restrictions in turn and accepting a model when the restriction gives a significant increase in χ^2 . Then we test the restrictions on error correlations in turn on the accepted structural model. The testing then continues on any remaining structural restrictions using the current error correlation model, and then again on the error correlation restrictions if necessary. The process stops if all more restricted models in both structural and error sense give significant increase in χ^2 . Models containing S1, E2 and E4 are not included as they are not homogeneous.

Table 2 shows the decision process for the test error correlation model. All structural restrictions are satisfied and the final model has 1 latent variable with equal correlations of test error across occasions of 0.55. Table 3 shows the decision process for the essay error correlation model. Here the restriction S4 is rejected in the context of E1 but accepted in the context of E3, the common essay error correlation being 0.21. The significant values given by the tests when used in this way cannot be given a rigid interpretation in probability terms as restrictions are tested sometimes more than once and a given model tested against more than one other model.

TABLE 2 Test error correlation model: decision process

Structural component	Goodness of fit	Difference	Significance Decision
No restriction on true scores	$\chi^2_3 = 1.44$		
Simplex restriction on true scores	$\chi^2_4 = 2.45$	$\chi^2_1 = 1.01$	> 0.05 accept
No change in true scores	$\chi^2_6 = 4.73$	$\chi^2_2 = 2.79$	> 0.05 accept
Measurement component (in context of S3) Equal error variances on all occasions (each test)	$\chi^2_{10} = 10.02$	$\chi^2_4 = 5.29$	> 0.05 accept
Equal error correlations across occasions	$\chi^2_{12} = 12.40$	$\chi^2_2 = 2.38$	> 0.05 accept
Zero error correlations across occasions	$\chi^2_{18} = 36.44$	$\chi^2_1 = 24.04$	< 0.001 reject

Table 3 Essay error correlation model : decision processes

Structural component	Goodness of fit	Difference	Significance(p)	Decision
No restriction on true scores	$X^2_3 = 7.V8$			
Simplex restrictions on true scores	$X^2_4 = 11.83$	$X^2_1 = 4.65$	< 0.001	reject
No change in true scores	$X^2_6 = 11.94$	$X^2_3 = 4.76$	< 0.01	reject
Measurement component (in context of S1)				
Equal error variances	$X^2_7 = 10.95$	$X^2_4 = 3.19$	> 0.05	accept
Equal error correlations (E5)	$X^2_9 = 11.49$	$X^1_2 = 1.46$	> 0.05	accept
Zero error correlations	$X^2_{10} = 34.94$	$X^2_1 = 23.45$	< 0.001	reject
Structural component (in context of E5)				
Simplex restriction on true scores	$X^2_{10} = 12.85$	$X^2_1 = 1.35$	> 0.05	accept
No change in true scores (S4)	$X^2_{12} = 13.08$	$X^2_2 = 0.23$	> 0.05	accept
Measurement component (in context of S4)				
Zero error correlations (as final model in Table 2)	$X^2_{13} = 36.44$	$X^2_1 = 23.36$	< 0.001	reject

Note that the initial "essay error correlation" has a bad fit to the data whereas the "test error correlation" model fits the data adequately in its most general form.

The final models are thus different for both forms. For the "test error correlation" model the variances of the test and essay errors are 0.34, 0.45 respectively and the reliability of tests and essays are 0.66, 0.56 respectively. The "essay error correlation" model has values 0.16, 0.56 for the variances of tests and essay errors and 0.84, 0.44 for the reliability of tests and essays (reliability is defined as $R = \sigma_x^2 / \sigma_X^2$ where $\sigma_X^2 = \sigma_x^2 + \sigma_e^2$ = variance of observed scores = 1 in our case due to use of correlation matrix, σ_x^2 = variance of true scores, σ_e^2 = variance of measurement error).

The estimates of error variances or equivalently of reliabilities are thus dependent on the initial modelling framework.

Werts et al justify their initial decision on the "essay error correlation" model by reference to cited research which includes evidence that good handwriting leads to higher essay scores regardless of content. It would be equally possible to defend the "test error correlation" model on the basis of subject specific reactions to the testing process. Indeed a whole day of a recent Symposium (4th International Symposium on Educational Testing, Antwerp, June 1980) was devoted to one possible mediating component, test anxiety.

Change in Attainment Over time : Exploratory and Confirmatory Analyses

LISREL is now used in the analysis of the change in reading attainment between ages 7 and 11 and between ages 7, 11 and 16 on data from the National Child Development Study (NCDS), Davie et al, (1970). This study followed up about 17,000 children born in one week of March 1958 at ages 7, 11 and 16. An obvious limitation on the use of LISREL to which attention is drawn by Joreskog and Sorbom (1977) is the requirement that the measures used actually measure the latent traits or hypothetical constructs. The analysis of the Werts et al data and Appendix 3 shows that assumptions which we make on the nature of the correlation between the errors of measurement can be viewed as part of the definition of the latent traits.

On the NCDS data this form of analysis requires that we have a number of reasonable measures at each occasion of the trait in question.

The following measures of reading attainment from the NCDS study are considered and listed here with a code.

	MEASURE	CODE	VARIABLE NUMBER
Age 7	Reading Test (Southgate)	RTST7	1
	Teacher Rating of reading ability	RTR 7	2
	Teacher rating of child's present standard on reading scheme	RSTD 7	3
Age 11	Reading Comprehension Test	RTST 11	4
	Teacher ratings of the Use of Books	U of B 11	5
	Teacher rating of Oral Ability	OTR 11	6
	Verbal component of General Ability Test	VGAT 11	6
Age 16	Reading Comprehension Test	RTST 16	-
	Teacher rating of English ability	ETR 16	-

Whilst we have three seemingly valid measures of reading attainment at age 7 only the first two of the 11 year measures have face validity, the Oral teacher rating and the verbal component of the General Ability test perhaps measuring different but related skills. The teacher rating of English at age 16 may also lack face validity. The same reading test was used at ages 11 and 16.

Desirable qualities in the data and their preliminary examination

The maximum likelihood estimation procedure of LISREL requires for consistency of estimates that the data are multinormally distributed. Under these conditions the variance-covariance matrix between variables represents all the information in the data as all moments above the second are zero. Multinormality can be tested by various methods. These are well reviewed by Bock (1975) and Gnanadesikan (1977) and further methods given by Cox and Small (1978) and Barnet and Lewis (1979).

Multinormality requires both that marginal distributions are normal and that relationships between all variables are linear, (though it is not implied by this). In practice data usually differ from this ideal situation, for instance, due to the relationships between variables being non linear or due to the variables

not being marginally normally distributed. In these cases variables may be transformed so that one or more of these conditions hold. (The references given above refer as much to these transformations as to the testing of multinormality, the two issues being closely related). It is also conceivable that linear relations occur between latent variables even when they do not occur between the indicators. However this requires that the relationship between the indicators and latent variables is non linear. This cannot be modelled in LISREL though it can be modelled in other programs (of McDonald, 1980 and Clogg, 1977)[1] and this allows the use of categorical variables as indicators of continuous latent variables. Alternatively, non-linear factor analysis (McDonald 1967) allows the factors or latent variables to be non-linearly related.

In the present data the aim was to use initial transformations to render the data as far as possible linear. The reading test at 11 years was scaled to have a normal distribution and the reading test at 7 years old was scaled to have a linear relationship to the 11 year score, the non-linear relationship for the raw data being interpreted as due to a ceiling effect on the 7 year scores. This scaling reduces but does not eliminate the skewness of the 7 year distribution. Scaling of the 7 year score to have a normal distribution would give a non-linear relationship between the scores at the two ages.

The teacher ratings, rated on a scale from 1 to 5 are viewed as categorisations of an interval scaled variable and so the distributional aspects are considered relevant. The distributions generally differed from normality in having a negative kurtosis (except Oral TR at 11 years) and a transformation to normality increased the more extreme values. Such a transformation also changes the relation to the test scores. Thus it is not possible to examine the effect of such a transformation whilst retaining the relationship between all the variables on the present data. However analyses are presented with the ratings scaled to normality (OSc) and unadjusted (O) in order to examine the sensitivity of the parameter estimates and the particular parameters which are freed to this transformation.

Table 4 gives the correlations between the variables for both variable sets. It is seen that the transformation to normality of the teacher ratings reduces most of the correlations particularly those involving the teacher rating of Use of Books and the teacher ratings of Oral ability at 11 years, increasing none of the correlations significantly.

1. Also Muthen and Dalquist (1980.

TABLE 4

MATRICES OF CORRELATIONS FOR SELECTED SUBSETS OF NCDS READING ATTAINMENT DATA

DATA SET

"O"	R TST 7	1.000					
	R TR 7	0.738	1.000				
	R STD 7	0.705	0.635	1.000			
	R TST 11	0.632	0.614	0.530	1.000		
	U OF B 11	0.607	0.601	0.511	0.680	1.000	
	O TR 11	0.534	0.542	0.409	0.618	0.707	1.000

		R TST 7	R TR 7	R STD 7	RTST 11	U OF B 11	O TR 11
"OSc"	R TST 7	1.000					
	R TR 7	0.742	1.000				
	R STD 7	0.622	0.599	1.000			
	R TST 11	0.632	0.616	0.461	1.000		
	U OF B 11	0.497	0.494	0.386	0.562	1.000	
	O TR 11	0.458	0.468	0.355	0.537	0.616	1.000

		R TST 7	R TR 7	R STD 7	RTST 11	U OF B 11	O TR 11
"V"	R TST 11	1.000					
	R TR 11	0.747	1,000				
	R STD 11	0.713	0.705	1.000			
	R TST 11	0.638	0.616	0.546	1.000		
	U OF B 11	0.618	0.60	0.534	0.684	1.000	
	V CAT 11	0.680	0.635	0.577	0.745	0.663	1.000

	R TST 11	RTR 11	R STD 11	RTST 11	U OF B 11	VCAT 11

"2" is as "O" except the last column is deleted

The linearity of the relations between variables was examined for sets both with transformed and non-transformed teacher ratings by examining the contribution of the squared term to the regression of each variable on the 11 year test score. Although all but 1 coefficient was significant (on a sample size 10681), the increase in R^2 due to fitting this term was always less than 0.02 being greatest for the scaled data for Oral Teaching Rating (0.009) and for the Verbal General Ability (0.008) and for the non-scaled data for Reading Standard teacher rating (0.013) and verbal general ability (0.008). Except for the Reading Standard teacher rating at 7 years the scaled teacher ratings had less linear relationships to the 11 year test score than the non-scaled ratings.

Further examination on unscaled data taking a random sample of 500 cases and fitting terms up to x^5 in the regression gave non-significant coefficients except in the regression of General Ability test at 11 on the reading test at 11 where the x^2 coefficient was significant at the 5% level. In a regression of teacher ratings on the General Ability test none of the higher order coefficients were significant. Overall then the degree of non-linearity in the data was not considered substantial though the lower correlations of many of the teacher ratings with the normally distributed 11 year test score when transformed suggest the prescence of some degree of non-linearity at least in the transformed data.

The face validity of the teacher ratings of Oral ability and of the Verbal component of the General Ability test is examined by runs of LISREL on the 7 and 11 year reading attainments with each of these indicators included separately (datasets "O" and "V") and also with none of them (dataset "2" - 2 indicators only at 11 years). If all the other indicators are suitable then the suitability of these indicators will be judged by the similarity of the parameter estimates for the data sets which include them and those which do not. These consist of the correlations between latent variables at different occasions, the correlations between errors of measurement and the variance of the errors or, equivalently, the reliability of the measures.

Finally the large sample size results in a significant lack of fit for models which would fit for smaller samples sizes and thus using the criterion of overall fit to select the appropriate model sometimes results in a large number of parameters being fitted. One way round this problem is to use a nominal sample size (say 1000) as the basis on which to compare the models. As well as giving the fit of each model to the data, the value of the sample size at which the model fits at the 0.05 level is also given and when this rises to above the nominal sample size this procedure would choose this as the final model.

In addition we are interested in the extent of change in the estimated correlation between the latent variables as we proceed through the model choice process. This is affected by the particular error correlations which are different from zero. Appendix6.4 gives the appropriate parameterisation of this problem in the LISREL formal in order that error correlations between occasions can be estimated. Figure 3 describes the model and Appendix6.5 briefly describes and justifies the model choice procedure used, that of freeing the parameter with the largest first order derivative. First, exploratory analyses of the 7 and 11 year data are presented. Then the 7, 11 and 16 year data is examined via models which utilise both exploratory and confirmatory approaches.

Exploratory Analyses of 7 and 11 year data

Table 5 gives the model selection process for each data set. ρ (= ψ_{12}) is the estimated correlation of the latent variables between occasions and $n_{0.05}$ is the sample size at which the solution reaches the 0.05 significance level.

In the column labelled "largest first order derivative","largest residual" are the variable numbers involved.

It can be seen that when a positive error correlation between occasions is fitted, the correlation between latent variables is reduced and vice versa when a positive error correlation between variables at the same occasion is fitted. (For example data set OSc models 1 to 2, 4 to 5). The opposite effect occures when negative error correlations are fitted. In 5 out 14 occasions the largest first derivative occurred between the same variables as did the largest residual.

The estimates of ρ for each of the nonscaled data sets, 0.821, 0.829, 0.822 are quite similar, and more similar than are found using a nominal sample size of 1000.

TABLE 5 PROCESS OF MODEL SELECTION ON THE FOUR DATA SETS FROM NCDS

Data Set	Order of model selection	Fit (X^2)	p value	$n_{0.05}$	$\hat{\rho}$	largest first order derivative	largest residual
"O"	1	X^2_8 = 492.0		337	0.819	5.6	1.4
	2	X^2_7 = 197.0		805	0.842	1.2	3.6
	3	X^2_6 = 26.0		5177	0.828	2.6	1.6
	4	X^2_5 = 15.1	0.01	7852	0.819	3.6	3.6
	5	X^2_4 = 5.4	0.25	18773	0.821		
"OSc"	1	X^2_8 = 938.6		177	0.819	5.6	5.6
	2	X^2_7 = 106.1		1420	0.845	3.4	3.4
	3	X^2_6 = 38.1		3533	0.852	6.2	6.3
	4	X^2_5 = 27.1		4334	0.850	5.2	6.3
	5	X^2_4 = 15.7	0.004	6457	0.845	5.1	6.3
	6	X^2_3 = 7.6	0.04	10978	0.830		
"V"	1	X^2_8 = 233.7		540	0.856	1.6	3.4
	2	X^2_7 = 180.0		638	0.850	5.6	3.4
	3	X^2_6 = 96.2		1066	0.843	1.2	3.4
	4	X^2_5 = 19.3	0.002	4681	0.827	2.6	2.6
	5	X^2_4 = 4.8	0.31	16090	0.829		
"2"	1	X^2_4 = 158.7		637	0.845	1.2	3.4
	2	X^2_3 = 5.3	0.15	15740	0.822		

Table 6 gives the estimated error correlations and Table 7 gives the estimates of reliability.

The nature of the error correlations between the measures at age 7 and the first two measures of age 11 are seen to be little affected by the third test at age 11. However, the scaling of the teacher ratings markedly affects the error structure, giving a larger proportion of error correlations between occasions.

Similarly the reliability estimates of all other measures is little affected by the inclusion or the choice of the third test at age 11. The scaling of the teacher ratings however reduces the reliability estimates of the teacher rating and also affects those of the tests.

TABLE 6 ESTIMATES OF CORRELATIONS BETWEEN ERRORS OF MEASUREMENTS FROM FINAL MODELS IN THE FOUR DATA SETS

"O"

R TST 7	1					
R TR 7	-0.45	1				
R STD 7	0	0	1			
R TST 11	0	0	0	1		
U OF B 11	0	0	0	0	1	
O TR 11	-0.07	0	-0.03	0	0.28	1

"OSc"

	1					
	0	1				
	0	0	1			
	0	0	-0.14	1		
	-0.04	0.07	0	0	1	
	0	0.06	0	0	0.21	1

"V"

R TST 7	1					
R TR 7	-0.36	1				
R STD 7	0	0	1			
R TST 11	0	0	0	1		
U OF B 11	0	0	0	0	1	
V GAT 11	0.04	0	0	0	0.23	1

"2"

	1					
	-0.45	1				
	0	0	1			
	0	0	0	1		
	0	0	0	0	1	

TABLE 7 RELIABILITY ESTIMATES FROM FINAL MODEL ON THE FOUR DATA SETS

DATA SET	R TST 7	R TR 7	RSTD 7	R TST 11	U OF B 11	OT R11	VGAT 11
"O"	.83	.79	.59	.71	.66	.55	-
"OSc"	.77	.72	.50	.74	.42	.38	-
"2"	.84	.80	.59	.70	.66		
"V"	.83	.79	.62	.74	.67		.79

These results give some reassurance that the inclusion of the third indicator has not markedly affected the latent variables at each occasion, though of the two extra indicators the General Ability test has slightly more effect than the Oral teacher rating on the other parameters in the model. They also suggest that each measure is a valid one.

The addition of extra indicators provides extra information which reduces the standard errors of the estimates and so under these conditions they should both be added to the model. The standard error of the correlation between the latent variables at 7 and 11 years is reduced from 0.006 to 0.005 by the addition of the oral teachers' rating.

Analysis of data on NCDS at three occasions

The exploratory analysis just considered is useful to the present analysis in two respects. Firstly it suggests that either or both of the Oral teacher rating and the General Ability test can be used as an indicator of the 11 year reading attainment and secondly it suggests that the teacher ratings taken at age 11 have correlated errors and that these also occur between the reading test at 7 and the teaching rating of reading at the same age. The full path model for the relations between the latent variable at each occasion but where the errors are concerned uncorrelated is shown in figure 4. The correlation coefficients ρ_{ij} and the conditional correlation coefficients $\rho_{ij/k}$ where $\rho_{ij/k}$ is the conditional correlation between the latent variables at occasions i, j given the value of occasion k are related to β_1, β_2 and β_3 This model (M3) is a saturated model for the structural aspects of the model as all possible parameters in the model are estimated. The likelihood ratio test of goodness of fit of this model gives $X^2_{17} = 1090.7$

Two alternatives are now possible, either to elaborate the model in an exploratory sense by freeing error covariances corresponding to the highest first order derivatives or some other criterion or to hypothesise some particular structure on the errors corresponding to one or more test specific factors. We have seen (Appendix 3) that a test specific factor is a reformulation of a set of non zero error covariances when 3 indicators have mutually correlated errors. It is more restrictive though more easily interpretable when more than 3 indicators have mutually correlated errors. One important restriction of models of this type is that no test specific factor for all the three or less indicators at a particular occasion can be hypothesised as this would not allow the relation between the latent variables at different occasions to be uniquely determined. However it is natural to chose one test specific factor for all tests and another for all teacher ratings, these being considered orthognal. This model, M5, described in figure 5 gives a $X^2_9 = 185.6$, a substantial improvement in fit. To what extent is the error variance of the indicators accounted for by this model? Table 8 gives the error variance of each indicator showing the proportion accounted for by the test specific factor.

TABLE 8. CONTRIBUTIONS OF TEST SPECIFIC FACTOR TO TOTAL ERROR VARIANCE

	Total error variance	Test factor contribution to overall variance	Remaining variance	Proportion of variance accounted for by Test factor
Tests				
RTST 7	0.177	0.006	0.171	0.04
RTST 11	0.258	0.130	0.128	0.51
RTST 16	0.254	0.076	0.178	0.30
VGAT 11	0.267	0.002	0.269	0.01
Teacher ratings				
RTR 7	0.321	0.138	0.183	0.43
RSTD 7	0.407	0.019	0.388	0.05
U OF B 11	0.383	0.011	0.372	0.03
RTR 16	0.403	0.009	0.304	0.02

Thus a large proportion of the error variation in the reading tests at ages 11 and 16 is accounted for by the test specific factor corresponding to the tests and a large proportion of the error variation in the teaching rating of reading at age 7 is accounted for by the test specific factor corresponding to the teacher ratings. The proportions for the other indicators are all low. The correlation between errors on the reading tests at ages 11 and 16 fitted by this model is 0.392 and corresponds to the artefacts introduced into the data by using the same reading test at the two ages.

Examining the first order derivatives from the full path model with uncorrelated errors (M3) it is this error covariance which has by far the highest value and freeing only this covariance gives model M4 with $X^2_{16} = 362.0$, again a substantial improvement, the estimate of the error correlation being 0.416. This is larger than the previous value as there are now no constraints on the test specific factor due to the other tests.

Examination of the loadings (λ_{ij}) for model M5 shows that at each age the ratio of the lowest to the highest is greater than 0.85 suggesting that each indicator is given a similar weight in determining the latent variable. At 7 years the test is most influential and at 16 years least influential.

Is the fit good enough?

Although the goodness of fit of each of these models is highly significant the sample size, 8189 cases, is very large. Bentler and Bonnet (1981) argue that models can be compared within and between studies by using an index of incremental fit which gives the relative improvement in the fitting criteria from one model to another on the same data. A suggested index is the normed fit index $\Omega = (F_k - F_1) / F_0$ where k, 1 are the compared

models and o is a suitable null model and F_i is the value of the fitting criterion for model i. A suitable null model in our case is taken as consisting of 1 common latent variable for all tests on all occasions. A possible alternative is the model M3.

Table 9 gives the set of models described earlier together with the simplex model described in the first section (model M2) and the 1 latent variable model M_1 and gives values of χ^2 for each model.

TABLE 9 GOODNESS OF FIT AND NORMED INCREMENTAL FIT INDICES FOR A RANGE OF MODELS

	MODEL (i)	GOODNESS OF FIT (X^2)	Ω
m_1	One latent variable	$X^2_{20} = 5230.0$	-
m_2	Simplex true scores	$X^2_{18} = 1253.5$	0.760
m_3	Full path model	$X^2_{17} = 1090.7$	0.791
m_4	M_3 with one non zero error correlation	$X^2_{16} = 263.0$	0.930
m_5	M_3 with two test specific factors	$X^2_9 = 185.6$	0.964
m_6	M_5 with four error correlations	$X^2_5 = 7.7$	0.999

For model m_5 the value, 0.964, for Ω is amongst the highest given in the examples given in Bentler and Bonnet and suggests that this model could be considered a final model.

However by successively freeing error covariances in the context of model m_5 using the maximum first order derivative as an indicator of which to be freed we arrive at a final model m_6 having four extra correlations between errors giving $X^2_5 = 7.71, p > 0.1$. The first error covariance to be freed is that between the reading test at 7 years and the Reading Standard rating at 7. This gives slightly higher first order derivative than that between the 7 year test and teacher rating of reading indicated by the high correlation between the errors in these variables in the earlier analysis. Other freed error correlations are between the Verbal General Ability test and the teacher ratings at 11 and 16 and the correlation between the reading standard rating at 7 and teacher rating of English at 16.

Effect of model choice on parameter estimates

Table 10 gives estimates of $\rho_{ij/k}$ and the reliability estimates of each indicator when the test specific factors have been included in the true score for models M3, M4, M5 and M6.

TABLE 10 ESTIMATES OF CORRELATIONS BETWEEN LATENT VARIABLES AND RELIABILITY OF INDICATORS FOR 3 MODELS

		M_3	M_4	M_5	M_6
$\hat{\rho}_{21}$		.848	.858	.867	.861
$\hat{\rho}_{32}$		.988	.956	.957	.959
$\hat{\rho}_{31}$		.777	.794	.794	.805
$\hat{\rho}_{32/1}$		.989	.881	.887	.880
$\hat{\rho}_{31/2}$		-.756	-.175	-.247	-.144
$\hat{\rho}_{12/3}$		.840	.555	.608	.530
Reliability	(RTST7)	.78	.78	.83	.78
"	(RTST11)	.78	.86	.88	.78
"	(RTST16)	.79	.85	.83	.82
"	(VGAT11)	.71	.74	.73	.86
"	(RTR7)	.72	.72	.82	.56
"	(RSTD7)	.63	.63	.61	.80
"	(U OF B11)	.61	.63	.63	.65
"	(ETR16)	.65	.69	.79	.71

The fitting of test specific factors is seen to affect all the parameters, especially the conditional correlations which have more reasonable values. The standard errors of the estimated correlations are of the order 0.02.

An alternative to the LISREL approach: Use of Instrumental Variables

The maximum likelihood estimation procedure of LISREL is equivalent to the least squares estimation procedure and gives optimal estimates only when all the variables are normally distributed. However, we have seen that there is a conflict between this requirement and that of linear relationships between the underlying variables. The procedure of instrumental variables, applied to a similar data set by Ecob and Goldstein (1981), avoids this problem by producing an unbiased estimate of the correlation between "true scores" on the tests say at ages 7 and 11, under certain conditions. Ecob and Goldstein argued that these generally hold when the 16 year score is regressed on the 11 year score for tests of reading and mathematics attainment. The reader is referred to Appendix 6 for details of the differences between the two estimation procedures.

It should be noted that all the models considered in this paper are conditional models in that the latent variables are related by the structural equation $\eta = \gamma\xi + \zeta$, and the parameter γ defines the expected value of η for given, unobserved, ξ. For the unconditional model, the structural part of the model is $\eta = \gamma\xi$ and this is equivalent fixing the covariance, ψ_{12}, in Appendix 4 in terms of ψ_{11}, ψ_{22} as $\psi_{12}^2 = \psi_{11}\,\psi_{12}$. The choice between conditional and unconditional models in the analysis of change is discussed by Bock (1975), Goldstein (1979a, b), and Plewis (1981). Taking the two teacher ratings at age 7 separately as instrumental variables we obtain the following estimates for the test score reliability at age 7 and of the correlation between test scores across ages 7, 11.

Instrumental Variable	Estimated Reliability of 7 year Test	Estimated correlation of test "true scores" at ages 7 and 11
Teacher rating of reading ability	0.78	0.81
Teacher rating of reading standard	0.76	0.84

The values for the correlation of true scores between ages and of the reliability of the 7 year test are comparable to the correlations between latent variables produced by the structural relations approach of LISREL.

The estimate of the standard error of the correlation is higher (0.01) using the instrumental variable approach than the values (average 0.005) using the structural relations approach thus reflecting the extra information used in the latter approach. Formal estimates of the regression coefficients of second occasion true score on first occasion true score and of its variance under a variety of structural equation models can be found in Joreskog and Sorbom (1974).

The results of the linear structural relations approach may be used to aid in the choice of appropriate instrumental variables: a correlation between errors on the teacher rating of reading and the reading test age 7 is indicated which would lead to a correlation of observed teacher rating with test score error which gives rise to an underestimate of the correlation of reading attainment across ages. As non-zero error correlations involving the reading standard rating are fitted this may be a more appropriate choice as instrumental variable.

Finally the models M5 and M6 at these occasions are compared with an instrumental variable analysis of Goldstein (1979a) in Table 11.

TABLE 11 COMPARISON OF LATENT STRUCTURE AND INSTRUMENTAL VARIABLE RESULTS ON READING ATTAINMENTS AT AGES 7, 11, 16

		Residual Variance
Model M5	$r_{16} = \underset{(0.025)}{1.084}\, r_{11} - \underset{(0.022)}{0.146}\, r_7$	0.06
Model M6	$r_{16} = \underset{(0.022)}{1.029}\, r_{11} - \underset{(0.018)}{0.080}\, r_7$	0.005
Instrumental variable method (IVM)	$r_{16} = \underset{(0.059)}{1.113}\, r_{11} - \underset{(0.051)}{0.147}\, r_7$	0.30

The lower residual in the latent structure models is partly due to the contribution of measurement errors in the 16 year test which is taken into account. The coefficients in M5 and IVM are within sampling error but M6 gives lower values for each coefficient.

Possible extensions to include more than one attainment and experimental or background variables

The structural equation modelling can be further extended to the more complex cases considered by Goldstein (1979a)of incorparating social class and examining relationships between reading and mathematics attainment at different ages. The models used here have been solely solely concerned with the (y, η) part of the LISREL model and Social Class would be included in the (x, ξ) part possibly measured with error using more than one indicator. An example of this form of analysis is given by Wheaton et al (1977). However, one needs to be aware of the assumption made in these LISREL models that the errors in the indicators are uncorrelated with the background variables. Results of Ecob (1979) suggest this is not true for this data. Hauser (1972) gives a model which allows for a direct effect of background variables on indicators.

The extra attainment can be modelled by considering more than one latent variable and specifying the loadings of some indicators on some latent variables as zero. An alternative method of examining the effects of Social Class on change in reading attainment is to consider the initial attainment measures as fixed variables also in the (x, ξ) part of the model. This loses the advantage of modelling test specific factors over occasions and we have no parameter corresponding to the correlation between latent variables over occasions. However the interpretation given to the effects of Social Class is improved. Joreskog and Sorbom (1977) give models similar to those in this section.

REFERENCES

V Barnett, & T Lewis (1978) Outliers in Statistical Data, New York, Wiley 365 pp

P M Bentler & P G Bonnet (1981) Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin 88, pp 588-606

H. Blok & W Saris (1981) Using longitudinal data to estimate reliability: a new look at the data of Werts, Breland, Grandy & Rock. (Internal Report).

D Bock (1975) Multivariate Statistical Methods in Behavioural Research, New York: McGraw-Hill.

C Clogg (1977) Unrestricted and restricted maximum likelihood latent structure analysis: a manual for users". University Park, Pennsylvania: Population Issues Research Office, Working Papers, 1977-9.

C Clogg (1981) New developments in Latent Structure Analysis. In D J Jackson and E F Borgatta (Ed) Factor Analysis and Measurement in Sociological Research.

H L Costner and R Schoenberg (1973) Diagnosing Indicator ills in multiple indicator models. In "Structural Equation Models in the Social Sciences". Eds: A Goldberger & D Duncan, Academic Press.

D R Cox & N J H Small, (1978) Testing Multivariate normality. Biometrika 65, pp 263-72.

R Davie, N Butler & H Goldstein (1970) From Birth to Seven: A report on the National Child Development Study.

R Ecob (1979) Discussion contribution to H Goldstein (1979) JRSS, A.142

R Ecob & H Goldstein (1981) Instrumental Variable methods for the estimation of test score reliability (submitted for publication).

R Gnanadesikan (1977) Methods for Statistical Data Analysis of Multivariate Observations. New York: Wiley, 311 pp.

H Goldstein (1979a) Some models for analysing data on Educational attainment (under discussion). JRSS A.142, pp 407-442

H Goldstein (1979b) The design and analysis of longitudinal Studies Academic Press, London.

R M Hauser (1972) Disaggregating a social-psychological model of Educational Attainment. Social Science Research. 1, 159-218.

K G Joreskog (1977) Structural equation models in the social sciences: Specification, estimation and testing in "Applications of Statistics". Ed. P R Krishnaiah, North Holland.

K G Joreskog and A S Goldberger (1975) Estimation of a model with multiple indicators and multiple causes of a single latent variable. JASA, Vol.7, No.351 pp 631-9.

K G Joreskog and D Sorbom (1974) Some regression estimates useful in the measurement of change. Research report 74-2, University of Uppsala.

K G Joreskog and D Sorbom (1977) Statistical Models and methods for analysis of longitudinal data. In "Latent variables in socio-economic models". Eds. D J Aigner and A S Goldberger, North Holland.

K G Joreskog and D Sorbom (1978) LISREL IV: Users Guide.

R P McDonald (1967) Non linear factor analysis. Psychometric monograph No 15.

R P McDonald (1980) A simple comprehensive model for the analysis of covariance structures. Some remarks on applications. Br.J. of Maths & Statist. Psychology 33

B Muthen & B Dalquist (1980) LADI-A. Latent analysis of dichotomous indicators - Version A, User's Guide, Department of Statistics, University of Uppsala.

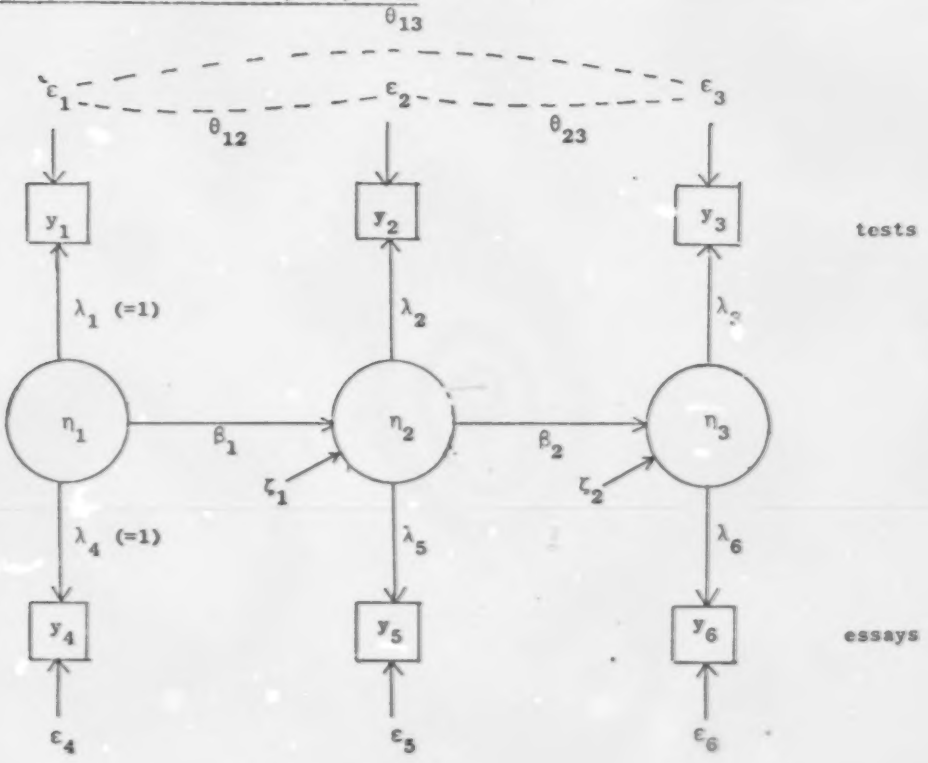
I Plewis (1981) Analysing Change: Using longitudinal data for the explanation and measurement of change in the social and behavioural sciences. Final Report to the SSRC.

D Sorbom (1975) Detection of correlated errors in longitudinal data. British Journal of Mathematical and Statistical Psychology. 28, pp 130-151.

D E Werts, H M Breland, J Grandy and D R Rock (1980) Using longitudinal data to estimate reliability in the presence of correlated measurement error. Educational & Psychological Measurement, 40, pp 19-29.

B Wheaton, B Muthen, D Alwin & G Summers (1977) Specification and Estimation of panel models incorporating reliability and stability parameters. In D R Heise (Ed) Sociological Methodology 1977.

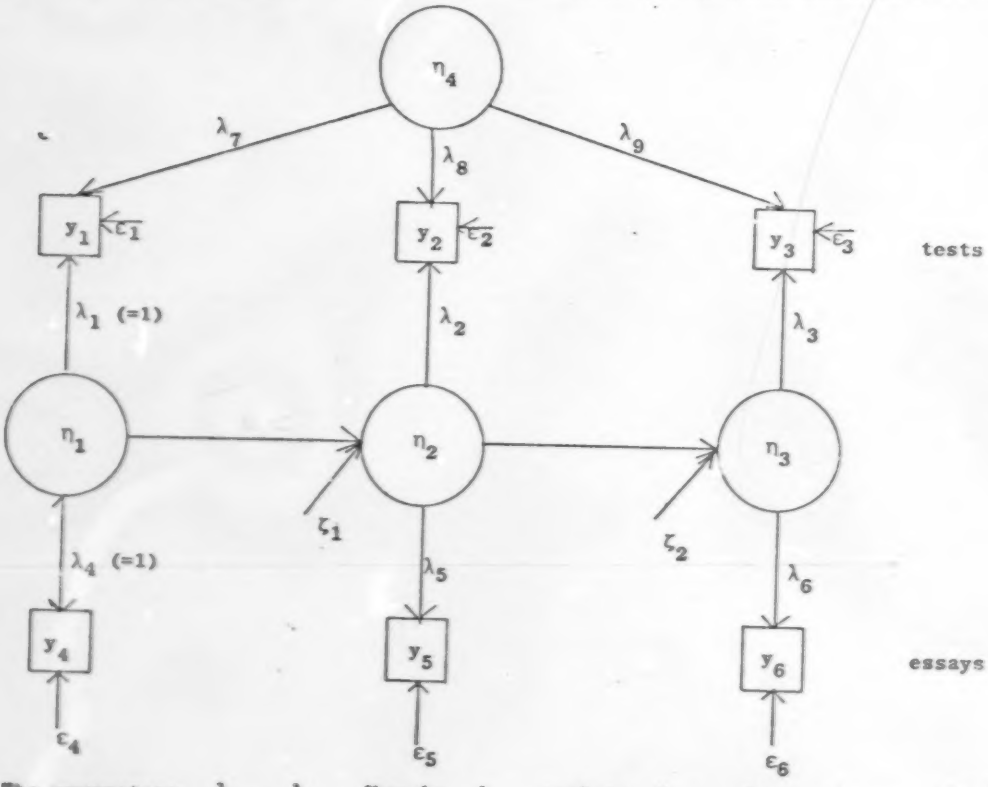
Figure 1. 2 Indicators of 1 latent variable (true score) at each of three occasions : model for Werts et al (1980) data



Simplex model for true scores; errors on one set of indicators are correlated.

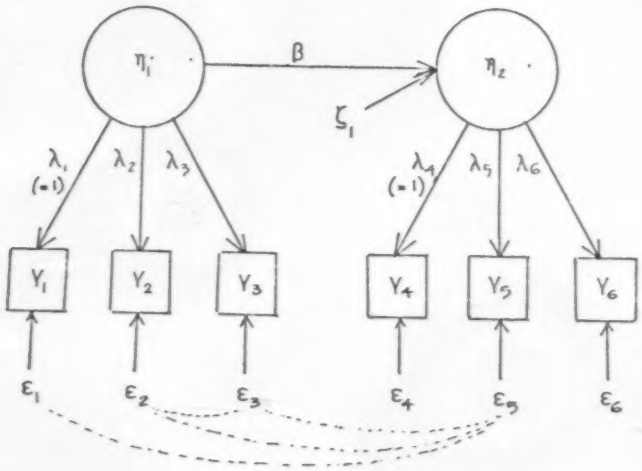
NOTE: The more general structural model has a coefficient β_3 relating η_3 and η_1 independently of η_2 giving the following matrix for $\underline{\beta}$, $\begin{bmatrix} 1 & 0 & 0 \\ -\beta_1 & 1 & 0 \\ -\beta_3 & -\beta_2 & 1 \end{bmatrix}$

Figure 2 Reformulation of error correlations in Figure 1 as a test specific factor



The parameters λ_8, λ_9, Var (η_4) replace θ_{12}, θ_{23}, θ_{13}, in figure 1. (see also Appendix 3).

Figure 3 A model with 2 occasions and 3 indicators at each occasion : 1 latent variable at each occasion



Errors $\varepsilon_1\varepsilon_5$, $\varepsilon_3\varepsilon_5$; $\varepsilon_2\varepsilon_3$; $\varepsilon_2\varepsilon_5$ are assumed correlated.

Figure 4 A model for the NCDS data on 3 occasions with one latent variable on each occasion with uncorrelated errors

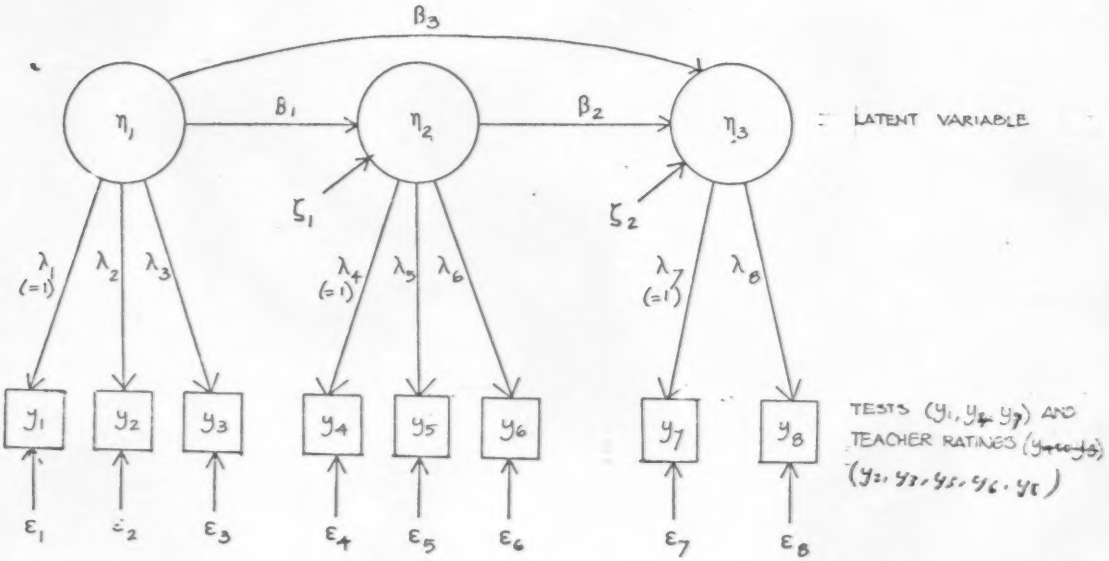
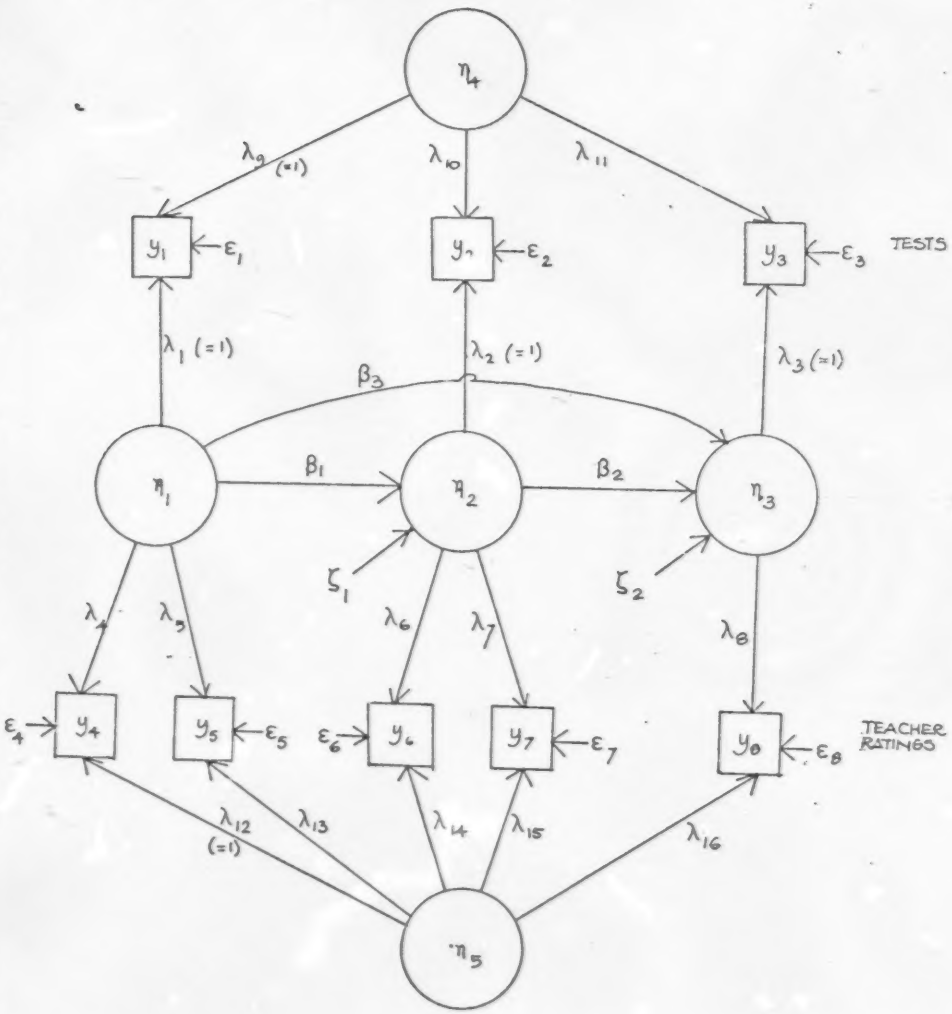


Figure 5 The 3 occasions model used on the NCDS data with additional test specific latent variables (model M5)



APPENDIX 6.1 LISREL SPECIFICATION OF WERTS ET AL MODEL

The simplex true score model used is formulated as $\underline{y} = \underline{\lambda}\,\underline{\eta} + \underline{\varepsilon}$ $\quad \underline{\beta}\,\underline{\eta} = \underline{\zeta}$

where $\underline{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ -\beta_1 & 1 & 0 \\ 0 & -\beta_2 & 1 \end{bmatrix}$ $\quad \underline{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$ $\quad \underline{\Psi} = \begin{bmatrix} \psi_{11} & & \\ 0 & \psi_{22} & \\ 0 & 0 & \psi_{33} \end{bmatrix}$

$$\underline{\lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ \lambda_4 & 0 & 0 \\ 0 & \lambda_5 & 0 \\ 0 & 0 & \lambda_6 \end{bmatrix} \quad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix} \quad \underline{\varphi} = \begin{bmatrix} \theta_{11} & & & & & \\ \theta_{12} & \theta_{22} & \text{SYMMETRIC} & & & \\ \theta_{13} & \theta_{23} & \theta_{33} & & & \\ 0 & 0 & 0 & \theta_{44} & & \\ 0 & 0 & 0 & 0 & \theta_{55} & \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} \end{bmatrix}$$

Two methods exist for fixing the scale of the latent variables at each occasion. One method is to fix the loadings of one indicator at each age. Here we adopt the alternative solution of fixing an indicator, λ_1 , for the first occasion latent variable, η_1 , and fixing β_1 , β_2 to fix the scales of η_2 , η_3 in terms of the scale of η_1 .

We then use the standardised solution which sets the variances of each latent variable to 1 to obtain the correlation between latent variables at each occasion. The error variance of indicator y_i is then equal to $(1 - \lambda_i^2)$ where λ_i is the loading on the latent variable.

Control card listing, parameter specifications, LISREL estimates and standardised solution follow for the most general "essay error correlation" Model.

APPENDIX 2 IDENTIFICATION PROPERTIES OF THE WERTS ET AL MODEL

We show here why the model with one set of non zero error correlations (between tests or between errors) is identified but not the model with two sets of non zero error correlations.

We use the convention described in Appendix 1 which fixes the coefficients in the relation between latent variables and the loading λ_i, of one indicator y_i, on η_i. Taking the essay error correlations to be zero we have $\theta_{45} = \theta_{46} = \theta_{56} = 0$ the situation in the earlier model.

We proceed to use the covariances between y_1, y_2 : y_1, y_3, and y_4, y_5 to obtain a value for $Var(\eta_1)$ and λ_4, λ_5 and see that it depends on θ_{45} being known.

We have

$$Cov(y_1 y_4) = \lambda_1 \lambda_4 \; Var(\eta_1)$$
$$Cov(y_1 y_5) = (\lambda_1 \eta_1 + \varepsilon_1)(\lambda_5 \eta_2 + \varepsilon_5)$$
$$Cov(y_4 y_5) = (\lambda_4 \eta_1 + \varepsilon_4)(\lambda_5 \eta_2 + \varepsilon_5)$$

We have also $\eta_2 = \eta_1 + \zeta_2$ and as $Cov(\eta_2, \zeta_2) = Cov(\eta_1, \zeta_2) = 0$ $Cov(\eta_2, \eta_1) = Var(\eta_1)$. Similarly, $Cov(\eta_i, \zeta_j) = 0$ FOR ALL i AND j and λ_i is fixed at 1.0

So we obtain
$$Cov(y_1, y_4) = \lambda_4 \; Var(\eta_1) \quad (1)$$
$$Cov(y_1, y_5) = \lambda_5 \; Var(\eta_1) \quad (2)$$
$$Cov(y_4, y_5) = \lambda_4 \lambda_5 \; Var(\eta_1) \quad (3)$$

Dividing equation 3 by the product of equations 1 and 2 we obtain $Var(\eta_1)$ and insertion in equations 1, 2 give λ_4, λ_5 respectively. If θ_{45} is not known then equation 3 becomes

$$Cov(y_4 y_5) = \lambda_4 \lambda_5 Var(\eta_1) + \theta_{45}$$

and the 3 equations cannot be solved. Moreover there are no other equations relating λ_1, λ_4 and λ_5.

Repeating this process with y_1, y_5, y_6 gives λ_6 and gives overidentification of $Var(\eta_1)$, λ_5 and $Cov(y_5, y_6) = \lambda_5 \lambda_6 \; Var(\eta_2)$ giving $Var(\eta_2)$.

λ_2, λ_3 are found by $Cov(y_1, y_4)$, $Cov(y_5, y_4)$ or by $Cov(y_2, y_5)$, $Cov(y_3, y_5)$ giving again overidenification, $Var(\eta_3)$ is found by $Cov(y_5 y_6) = \lambda_5 \lambda_6 Var(\eta_3)$ and finally $\theta_{12}, \theta_{13}, \theta_{23}$ are found by $Cov(y_1 y_3)$, $Cov(y_2 y_3)$, $Cov(y_1 y_2)$.

4 parameters, $Var(\eta_1)$, $\lambda_6, \lambda_2, \lambda_3$ are obtained by either of 2 equations giving 4 degrees of freedom for testing the model.

Let us consider a model where the indicator which has correlated errors instead has loadings on a test specific factor (η_4), the errors being then uncorrelated. The other indicator has no loadings on this factor.

We then have

$$\underline{y} = \underline{\lambda}_y' \, \underline{\eta}' + \underline{\varepsilon} \quad , \qquad \underline{\beta}' \underline{\eta}' = \underline{\zeta}'$$

where

$$\underline{\lambda}_y' = \begin{bmatrix} \lambda_1 & 0 & 0 & \lambda_7 \\ 0 & \lambda_2 & 0 & \lambda_8 \\ 0 & 0 & \lambda_3 & \lambda_9 \\ \lambda_4 & 0 & 0 & 0 \\ 0 & \lambda_5 & 0 & 0 \\ 0 & 0 & \lambda_6 & 0 \end{bmatrix} \qquad \beta' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \Psi = \begin{bmatrix} \psi_{11} & & & \\ & \psi_{22} & & \\ & & \psi_{33} & \\ & & & 0 \end{bmatrix}$$

$$\underline{\eta}' = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix} \quad , \quad \underline{\varepsilon}, \ \underline{\eta}_2, \ \underline{y} \ \text{AS IN APPENDIX 1} \ \lambda_1, \lambda_4 \ \text{ARE FIXED}$$

λ_7 is fixed and the identification process is identical to that given in Appendix 2 apart from Cov (y_1, y_4), Cov (y_1, y_5), Cov (y_2, y_5)

Now Cov ($y_1 y_4$) = $\lambda_1 \lambda_4$ Var (η_1) + $\lambda_7 \lambda_8$ Var (η_4)

and as $\lambda_1 \lambda_4$, Var (η_1), λ_7 are known we obtain λ_8 Var (η_4).

Similarly Cov (y_1, y_5), Cov (y_2, y_5) give λ_5Var (η_4), $\lambda_8 \lambda_9$Var (η_4) and by a similar process to the initial identification in the earlier model we obtain values for each of these quantities.

So, for the case where 3 indicators are correlated this is identical to the existence of a test specific factor for these indicators, the loadings being the same when the covariances are equal. Note that an error correlation between only 2 indicators does not imply the existence of a test specific factor as the two parameters involved cannot be determined by the one covariance and where more than three indicators are present overidentification of this factor will result.

As the model with error correlations between both sets of indicators is unidentified so also is the equivalent model with a test specific variable for each type of indicator.

APPENDIX 4 LISREL SPECIFICATION OF 2 OCCASION MODEL WITH 3 INDICATORS AND 1 LATENT VARIABLE AT EACH OCCASION

The model following allows for correlation of errors between occasions and has nq x variable, and is identical to that in Sorbom (1975) and to the confirmatory factor analysis model, Example 5 in the LISREL Manual.

$$\underline{y} = \underline{\lambda}_y \underline{\eta} + \underline{\varepsilon}, \qquad \lambda_y = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & \lambda_4 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \end{bmatrix}$$

$$\underline{\eta} = \underline{\zeta}$$

$$\underline{\Psi} = \begin{bmatrix} \psi_{11} & \\ \psi_{21} & \psi_{22} \end{bmatrix}$$

λ_1, λ_4 FIXED (AT 1)

is the correlation of the latent variable between occasions.

It is important to note that the LISREL User's Guide (1978) gives the following model for change in ability between occasions (Section III.3 and Example 7).

$$\underline{y} = \underline{\lambda}_y \underline{\eta} + \underline{\varepsilon} \qquad \eta = \gamma\xi + \zeta$$

$$\underline{x} = \underline{\lambda}_x \underline{\xi} + \underline{\delta}$$

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \quad \underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad \underline{\lambda}_y = \begin{bmatrix} 1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \quad \underline{\lambda}_x = \begin{bmatrix} 1 \\ \lambda_4 \\ \lambda_5 \end{bmatrix}$$

η, ξ, ARE LATENT VARIABLES

However, this only allows error correlations within occasions to be fitted.

Control card listing parameter specifications and fitted estimates follow for the final model for data set "0".

APPENDIX 5 MODEL CHOICE PROCEDURE IN AN EXPLORATORY SITUATION

The literature of Joreskog and Sorbom, is confusing on the recommended procedure for model choice in this situation. Joreskog and Sorbom (1977) state

> "In a more exploratory situation the χ^2 goodness-of-fit values can be used as follows. If a value of χ^2 is obtained which is large compared to the numbers of degrees of freedom, the fit may be examined by an inspection of the residuals, i.e. the discrepancies between the observed and the reproduced variances and covariances. The examination, in conjunction with subject-matter considerations, may suggest ways to relax the model somewhat by introducing more parameters. The new model usually yields a smaller χ^2. A large drop in χ^2, compared to the difference in degrees of freedom, supports the changes made. On the other hand, a drop in χ^2 which is close to the difference in number of degrees of freedom indicates that the improvement in fit is obtained by capitalizing on chance."

However, in Joreskog (1977) we find the same paragraph except that the underlined words are replaced by "by an inspection of the magnitude of the first derivatives of F with respect to the fixed parameters." In fact earlier literature, Costner and Schoenberg (1973), Sorbom (1975) points to deficiencies in the analysis of residuals using simulated 2 occasion data of the type used in our second example which include non zero correlations between errors within occasion (Costner and Schoenberg) and within and between occasions (Sorbom), this shows that the iterative procedure which frees the largest residual at each stage results in the incorrect parameters being freed. Costner and Schoenberg find that the correct model is found by an analysis of the set of submodels which exclude at least one indicator at each occasion, and Sorbom finds that an analysis of the first order derivatives gives the correct model. The latter method is used here as it is much more economical. However, as Sorbom mentions, the freeing of the parameter with largest first order derivatives will not give the largest decrease in χ^2 as another parameter with greater change in its value between models could theoretically give a larger decrease in χ^2. Some idea of whether the models found are the correct ones can be given by comparing the order of freeing the first derivatives with the ordering of the freed correlations in each data set.

In "O" none out of a possible 6 order changes occur,
in "OSc" 1 out of a possible 10 and in "V" 4 out of a possible 6 occur.
This gives some limited confidence that the best model with the given number of freed parameters has been found. However, imposing the final model of "O" on "OSc" gives a lower χ^2 (χ^2 = 15.2) then the equivalent model in fact chosen (χ^2 = 15.7). Only one of the parameters freed (5, 6) was the same in each case.

APPENDIX 6 : THE LINEAR STRUCTURAL RELATIONS (LSR) AND INSTRUMENTAL VARIABLE (IV) ESTIMATORS COMPARED

	IV	LSR
Estimation method	Least Squares	Maximum likelihood
Condition for consistency estimates of regression (or correlation) coefficient	IV's are not correlated with errors of measurement or disturbances	Correct model is chosen, and observed variables are jointly normally distributed
Effect of correlations of measurement errors between occasions on the regression coefficient.	None	None, if incorporated into the model.
Effects of correlations of measurement errors within occasions on the regression coefficient	Inconsistency of this leads to a correlation between the observed instrumental variable and test score measurement error.	None, if incorporated into the model
Action possible when there are non linear relations between normally distributed variables	Transform independent variable to a linear relation with dependent variable	No action possible to give consistent estimates
Efficiency of estimates	Dependent on the correlation of instrumental variable(s) with independent variable - generally suboptimal	efficient if variables are normally distributed
Effect on variance of regression coefficient of increase in number of tests at either occasion	If all used as instrumental variables, variance generally decreases (See Cobb & Goldstein, 1981)	Variance generally decreases (See Joreskog and Sorbom, 1974)

APPENDIX 7

ESTIMATING THE INCONSISTENCY OF INSTRUMENTAL VARIABLES ESTIMATES IN THE CASE OF CONGENERIC VARIABLES WITH CORRELATED ERRORS

By Russell Ecob

The three variables, the independent, dependent and instrumental variables are represented as congeneric variables with correlated errors, this being known as a reformulation of the test specific latent variable (see Appendix 6). The dependent variable, however, contains two error components, one being a disturbance term in the regression on the independent variable which is assumed to be uncorrelated with the error of measurement of the instrumental variable.

Let x_1 , x_2 , x_3 be the observed values of the independent and instrumental variables respectively.

Then we have

$$x_1 = \beta_1 t + e_1$$

$$x_2 = \beta_2 t + e_2 + u$$

$$x_3 = \beta_3 t + e_3$$

where $\Sigma e_i t = 0, \; i = 1,3$, $\Sigma e_3 u = 0$ and $\Sigma t u = 0$

and $E(e_i) = 0$, $E(u) = 0$

The true scores , t_1 , t_2 at occasions 1 and 2 are given by

$$t_2 = \beta_2 t + u$$

$$t_1 = \beta_1 t$$

giving the following relation between true scores;

$$t_2 = \frac{\beta_2}{\beta_1} t_1 + u$$

or $t_2 = \beta t_1 + u$ where $\beta = \beta_2 / \beta_1$ is the regression coefficient in the relation between the true scores at the two occasions.

Let us denote the correlation between errors e_i, e_j on variables x_i and x_j by ρ_{ij} and let the reliability of x_i be R_i .

The Instrumental Variables estimate of the regression coefficient of x_{2on} on x_1 is

$$\beta_{IV} = \frac{\Sigma x_2 x_3}{\Sigma x_1 x_3} = \frac{\Sigma (\beta_2 t + e_2 + u)(\beta_3 t + e_3)}{\Sigma (\beta_1 t + e_1)(\beta_3 t + e_3)}$$

and as $R_i = \frac{Var(\beta_i t)}{Var(x_i)} = \frac{\beta_i^2}{\beta_i^2 + Var(e_i)}$ after some simple algebra we obtain

$$\beta_{IV} = \frac{\beta_2\left(1 + \rho_{23}\sqrt{\frac{1-R_2}{R_2}\cdot\frac{1-R_3}{R_3}}\right)}{\beta_1\left(1 + \rho_{13}\sqrt{\frac{1-R_1}{R_1}\cdot\frac{1-R_3}{R_3}}\right)} \quad \ldots (1)$$

and the relative inconsistency, $I = \frac{\beta_{IV} - \beta}{\beta}$ given by $\frac{A_2 - A_1}{1 + A_1}$

where $A_i = \rho_{i3}\sqrt{\frac{1-R_i}{R_i}\cdot\frac{1-R_3}{R_3}}$

When x_1, x_2 have the same reliability, R

then $I = \frac{\beta_{IV} - \beta}{\beta} = \sqrt{\frac{1-R_3}{R_3}\cdot\frac{1-R}{R}}\left(\rho_{23} - \rho_{13}\right)\left[1 + \rho_{13}\sqrt{\frac{1-R}{R}\cdot\frac{1-R_3}{R_3}}\right]^{-1}$ (2)

$$\leq \sqrt{\frac{1-R_3}{R_3}\cdot\frac{1-R}{R}}\left(\rho_{23} - \rho_{13}\right)$$

For this particular model the inconsistency of the instrumental variables estimator is given in terms of the unknown correlation of indicator errors and reliabilities. However this model is similar to the structural relations model, where an extra indicator is required for the dependent variable, the true indicators of the independent variable being x_1 and x_3.

Two of the instrumental variables used in Appendix 5b, the verbal component of the General Ability test and the teacher rating of oral ability are used simultaneously as indicators of the 11 year reading attainment. Though the instrumental variables analysis uses each of these indicators separately as instrumental variables, the comparability of estimates of β formed by correcting the separate instrumental variable estimates using the error correlations and reliabilities estimated from the structural equations model will provide an indication of the consistency of the two approaches with each other. It is known also that fitting non zero correlations between errors of particular

indicators will affect other fitted error correlations in the model. Thus other indicators may obscure a non-zero correlation between two particular indicators. The effect of this is examined by forcing particular error correlations to be freed early in the fitting process.

The following is a list of the indicators used;

1) Test of reading at age 11

2) Test of reading at age 16

3) Oral teacher rating at age 11 (IV1 - 1st instrumental variable)

4) Verbal component of General Ability Test at 11 (IV 2)

5) English teacher ratings at age 16

As the analysis in Appendix 6 suggested, the largest error correlation by far is that between the two tests (1,2). Freeing the error correlations in terms of their highest first order derivatives freed (1, 3) and then (1, 4) this giving an acceptable fit to the data ($X_2^2 = 2.6$) and a fitting first (1, 4) gave again (1, 3) as the next

freed parameter producing the same model (Model 1).

Alternative models were obtained by fitting either (2, 4) or (2, 3) after (1, 2) giving the freed error correlations for adequately fitting models for Model 2 between (1, 2), (2, 4) and (1, 3) and for Model 3 between (1, 2) (2, 3) and (4, 5).

The estimates of the parameters for the different models is given in the table / below together with estimates of the inconsistency, I. Estimates of are then obtained using the instrumental variable estimator of 0.989 and 1.014 of IV1 and IV2 given in Appendix 5b.

		Model 1	Model 2	Model 3
Error Correlations	1,2	.102	.091	.062
	1,3	041	.039	-
	1,4	-.017	-	-
	2,3	-	-	-.037
	2,4	-	.016	-
	2,5	-	-	-.016
Reliability Estimates	Test	0.73	0.74	0.78
	IV1(3)	0.70	0.70	0.75
	IV2(4)	0.62	0.63	0.59
I IV1		-0.016	-0.015	-0.011
$\hat{\beta}$		1.005	1.004	1.006
I IV2		0.008	0.008	0.000
$\hat{\beta}$		1.006	1.006	1.014

The corrected estimates of β corresponding to the different instrumental variable estimators are very similar in Models 1 and 2 but differ more in Model 3. The Models 1, 2 also produce similar estimates of β. The estimates of correlations between the underlying variables at each age were found to be 0.981 using the corrected instrumental variables estimator using estimates of error variance from the structural equations analysis and 0.976 directly from the structural equations analysis.

APPENDIX 8

Missing data in the NCDS: the use and evaluation of a method of Beale & Little for estimating partially missing responses

By Russell Ecob

The Problem

The NCDS study was by most standards remarkable for the high follow up rate of respondents at earlier stages, 87% of the original birth cohort providing at least partial responses at 16 years, 91% at 11 years and at 7 years. Were those who did not respond at later stages different in any ways from the overall sample, particularly in ways which would affect their 16 year score? Goldstein in Fogelman (1976) addressed himself to this question and found a tendency for non respondents at 16 years who responded at earlier stages to come from disadvantaged groups at ages 7 and 11 including illegitimate children, those who received special education and those exhibiting "anti-social" behaviour in school, these categories providing 3 - 6% of the children in the survey.

In addition when those children with 16 year data were compared with those without, biases were found in county of residence, type of accommodation, the direction of bias being such that the proportion of children in categories associated with lower school attainment is underestimated when only children with some data at 16 years are considered. The exception to this rule was an over-representation of children from small families among those with no data at 16 years.

Goldstein also carried out analyses of the change in test scores between 7 and 11 years for those with data at these ages and with and without data at 16 years. The analyses considered the 7 year score as a covariate and either considered the response contrast between those with some data at 16 years versus those without or considered these categories separately in the latter case including Social Class or number of children in the household as a factor.

Whilst the response/non response contrast was significant for the mathematics attainment no difference in regression coefficient of 11 year on 7 year score for either mathematics or reading was found between those with some data compared to those with no data at 16 years, though a difference between non manual and Social Class V children was increased for reading and mathematics by 3 and 2% respectively. Extrapolating the response-non-response contrast for mathematics to 16 years score gave a bias of 0.05 years of attainment when estimates using the available data are used. An explanation for the significant response/non-response contrast for mathematics attainment not being shown up in a difference between regression coefficients in the "response at 16 years" group and the total group is the small percentage, 9%, who had no response at 16 years but who responded at 7 and 11 years.

We now ask how relevant such an analysis is to the problem of non-response in more complex analyses of the NCDS data which examine relationships between a number of variables at each age. Examples are the analysis of Goldstein (1979) and the instrumental variable estimation described in Chapter 2.2 . Both these analyses require data which have responses on both attainment tests, teacher ratings and social class at each of the three ages 7, 11 and 16. This gives sample sizes down to 5100 cases, less than a third of the overall sample and much less than the 8900 cases examined in Fogelman (1976) with relevant responses at 16 years.

All techniques which estimate the non-response bias need to make the assumption that, given the information recorded, a non-respondent on a particular variable is equivalent to a respondent. This is called the "missing at random" assumption by Marini, Olsen and Rubin (1980), It can be seen that any description of differences for instance, between children with and without 16 year data are restricted to the variables measured at earlier ages. Moreover the analysis which examined response bias at 16 years included only those children who were observed on the variables included in the analysis at ages 7 and 11 and even where only attainment tests were considered, excluded 47% of the total sample. The assumption was made that these 53% were the same as the rest of the sample in

the scores at 7 and 11 years.

Before describing an analysis which conceptualises missing values more generally we present an analysis which examines the difference between "16 year respondents" and non respondents where the "16 year respondents " group have responses on both reading and mathematics attainments at 16 and both groups have responses on these attainments at 7 and 11 years in addition to a measure of Social Class at 7 years. This analysis differs from that of Goldstein reported earlier in that out of 12864 having the responses at 7 and 11 years only 9420 (72.7%) have responses at 16 years on both attainments (this sample being similar to that in Goldstein (1979), Tables 2, 4).

Table *1* gives the regression coefficients at 11 year on 7 year scores for both attainments and the fitted constants corresponding to social class at 7 years when those with and without 16 year data are included.

TABLE *1* REGRESSION OF 11 YEAR ON 7 YEAR READING AND MATHEMATICS ATTAINMENTS SEPARATELY WITH AND WITHOUT 16 YEAR SCORES

			11 on 7 Yr Regression Coefficients	Social Class Fitted Constants		
Reading test	All data	12864	0.61	0.18	-0.08	-0.25
	16 year respondents only	9420	0.60	0.19	-0.09	-0.25
Mathematics test	All data	12864	0.55	0.24	-0.10	-0.37
	16 year respondents only	9420	0.54	0.25	-0.12	-0.40

Both 7 and 11 year tests are standarised so the regression coefficients are partial correlation coefficients when Social Class (in 3 categories) is controlled for. The regression coefficients differ by 1-2% in the two groups and the constants fitted to social class categories differ by 7-10%. These differences are much larger than those given in Fogelman (1976) and raise the question whether larger differences would be found if more data were excluded.

A more general conception of missing values

We now consider a missing value pattern. A particular pattern is obtained by the set of variables on which the values are missing and given a set of variables of interest a variety of patterns will occur. For instance in the previous analysis just two patterns were considered, complete information on all selected variables versus missing information on one or more 16 year attainments only. Given a selection of n variables the total number of possible missing value patterns is 2^n though not all of these will always occur. In general the more variables that are considered the more likely the "missing at random" assumption is to hold as differences on other variables for the cases which are missing on a given variable are found.

The Beale and Little method for estimating missing values

Beale and Little (1975) examine methods for estimating missing values which for any missing value finds an estimate by the regression on the variables having known values. The estimates are in turn taken as known values for the estimation of further missing values. The process is allowed to iterate until convergence occurs. Six methods are compared, 5 of which are maximum likelihood in some sense, 1 which uses only ordinary least squares estimation and 3 of which use a combination of both methods where firstly missing independent variables are fitted by modified maximum likelihood using the independent variables only and then a dependent variable is fitted by weighted least squares on the estimated values. The method used in this paper is Method 6, one of the last 3 methods where weights are estimated from the data. A more straightforward method using maximum likelihood estimation which does not require iteration is given by Marini, Olsen and Rubin (1980) which requires that the missing values

be nested so that no cases occur which both have variable x missing and variable y present and vice versa. This is considered likely to apply to longitudinal data where there is gradual exodus from the study but does not occur in the NCDS study and so this method is not examined further, though non-nested data can be considered as lying between two nested bounds by excluding some of the values. In addition the method of Marini, Olsen and Rubin requires normality of distributions of variables for the desirable properties of the estimates to be shown, this not being necessary for the least squares method of Beale and Little (1975).

Examination of the Beale and Little method in NCDS data

The Beale and Little method produces consistent estimates of the first and second moments of all variables in a dataset if the missing values are missing at random; irrespective of the distributional aspects of the data.

We now focus on the effect of using different sets of ancilliary variables for the estimation of relationships between certain 'key' variables. The fact that the relation of 7 to 11 year reading and mathematics tests differs for the respondents and non-respondents at 16 years implies that given this set of variables the 16 year test data are not 'missing at random'. So for the missing 16 year data using only 7 and 11 year test data to derive values for 16 year data will give biased regression coefficients of 16 on 11 year data. However

using additional variables, social class and teacher rating measured at each age, we have extra information on individuals with one or more test scores missing at age 16 and in some cases even data at this age.

It is likely that for the cases having observations on these extra variables this information can be used to reduce the inconsistency in the estimation of parameters pertaining to the missing values. Two factors are likely to limit the usefulness of this method. Firstly the dependancy of the"missingness"on variables which are not included in the ancilliary variable set. For instance the non-respondents at 16 years were found to be over represented in small families. If this variable has an effect on the 16 year scores after controlling for the ancilliary variables then estimates will be inconsistant. Secondly the ancilliary variables are often themselves missing in many of the observations with missing values on the 16 year scores. For these cases the ancilliary variables can only be used indirectly through the preliminary estimation of the missing values of the ancilliary variables through other variables.

However, the use of social class and also measures of attainment at 16 years as ancilliary variables is expected to control to some extent for the dependency of the"missingness" of the test data on other factors.

The following analysis examines the effect of the internolation of missing values using various subsets of ancilliary variables on various estimates of the relation of 16 year to 11 year and of 11 year to 7 year tests of reading.
These values are compared to those obtained by using only cases with values on all these variables and using cases with values on all variables involved directly in estimation of the parameters associated with 16 year and 11 year tests and the relationship between them.

The data and the missing value patterns

10 variables were included, these being tests of reading attainment and teacher ratings of reading and Social Class at each of these ages 7, 11 and 16 and also a test of General Ability at 11 years. The key variables were the reading tests at ages 11 and 16 and the teacher rating of reading at age 11, other variables being ancilliary. Excluding cases with no observation on any of these variables left 17070 cases. 231 out of the 1024 possible missing value patterns were found and of these 5 each accounted for greater than 3% of the data and a further 12 for between 1 and 3% of the data. These are given in Table 2 where 0 denotes the value of this variable being missing and 1 present.

TABLE 2 Missing value patterns accounting for greater than 1% of the data (all lists and teacher ratings are of reading attainment

	test 7	tr 7	sc 7	test 11	general ability 11	tr 11	sc 11	test 16	tr 16	sc 16	frequency	%
> 3%	1	1	1	1	1	1	1	1	1	1	6114	35.8
	1	1	1	1	1	1	1	1	1	0	1795	10.5
	1	1	1	1	1	1	1	0	0	1	923	5.4
	1	1	1	1	1	1	1	0	0	0	1378	8.1
	1	1	1	0	0	0	0	0	1	0	577	3.4
< 3%	1	1	1	1	1	1	1	1	0	1	205	1.2
but ≥ 1%	1	1	1	1	1	1	1	0	1	1	256	1.5
	1	1	1	1	1	1	0	1	1	1	437	2.6
	1	1	1	1	1	1	0	1	1	0	255	1.5
	1	1	1	1	1	1	0	0	0	0	195	1.1
	1	1	1	0	0	0	1	1	1	1	334	2.0
	1	1	1	0	0	0	0	1	1	1	301	1.8
	1	1	1	0	0	0	0	1	1	0	176	1.0
	1	1	0	1	1	1	1	1	1	1	216	1.3
	0	0	0	1	1	1	1	1	1	1	341	2.0
	0	0	0	0	0	0	0	1	1	1	236	1.4
	0	0	0	0	0	0	0	1	1	0	166	1.0
											13827	81.6
											17070	100

Only 50.5% of the data conform to patterns where all data at a given age are either present or absent and so much information on missing values is generally provided by other data collected at the same occasion. 28% of the data has information missing only at 16 years and 19.8% had observations on teacher ratings and or on the general ability test which were not on the reading test of that age or vice versa.

The Beale and Little program was run on a random sample of 919 cases, this being near the maximum sample size within computer memory limitation. It was found that a reordering of the data on the three test variables to bring similar missing value patterns together in the ordering improved the performance by a factor of 0.3 and so this was done for all runs. Estimates were obtained for the full data of the mean test scores on the reading test at 7, 11 and 16 years and the mean values of the teacher ratings of reading at 7 and 11 years. Also obtained were the ordinary least squares and instrumental variable regression coefficient of 16 year on 11 year and 11 year on 7 year reading tests and the derived reliability estimates of the reading tests at 11 and 7 years as outlined in Chapter 2.2. The instrumental variables used in the two regressions were the teacher ratings of reading at 11 years and 7 years respectively.

Ten runs were made. The first two, numbers 1, 1a, did not use the missing values program and gave estimates respectively for the data sets with complete values of all key and ancilliary variables and for the data sets with complete values of the key variables only. These are called the "complete case" and "key" datasets respectively. Datasets 2-9 gave estimates using the missing value program using different values of ancilliary variables generated according to the following pattern.

Social Class	+		-	
Ancilliary attainment measure	+	-	+	-
11 year data on ancilliary variables	2	4,5	3	/
No 16 year data on ancilliary variables	6	8,9	7	/

Here + denotes the inclusion of the particular ancilliary variable and - the exclusion. The ancilliary attainment measure includes the teacher ratings and also the general ability test at 11 years. Runs 4 and 5 differ by excluding in run 5 the teacher rating at 11 years. Similarly for runs 8,9. Runs 6-9 exclude all 16 year variables fr· the data and so exclude respectively social class and teacher ratin,; Social cla ; teacher rating at 16 years in runs 6; 8, 9; and 7. Run 2 includes all ancilliary variables.

The analysis of the means and of the regression coefficients is considered separately.

The means of the three key variables and the 7 year teacher rating are given here for each of runs 1, 1a, 2.

Run No.	Sample size	tests 7 year	tests 11 year	tests 16 year	teacher ratings 7 year	teacher ratings 11 year
1 (Complete case)	2.95	28.8	0.11	0.11	2.71	2.77
1a (wrong)	500	/	0.09	0.09	/	2.80
2 (all available)	919	27.5	0.04	0.07	2.86	2.86

All other runs 3-9 gave estimates close to run 2.

Expressing the changes between runs relative to the estimated standard deviation in the estimated dataset 2, these were in percentages.

percentage change (in current)

Run	tests 7 year	11 year	16 year	teacher rating 7 year	11 year
1 v 1a	/	2.8	1.4	/	3.6
1a v 2	/	5.5	9.3	/	7.2
1 v 2	3.1	8.3	10.7	21.2	10.8
2 v 3-9	0.5	0.2	1	0	0

The values in the row 2 v 3-9 are the maximum difference between any of runs 3-9 and run 2 and shows that even using only the 11 year attainments measures as ancilliary variables (run 7) gives values very close to the estimates obtained using all ancilliary variables (a maximum difference of 0.006, in the 16 year test, ever found). In contrast the estimation using ancilliary data produced estimates substantially different (max 9.3% on 16 year test) from the "key" dataset and this in turn produced estimates differing in the same direction from the complete case datasets (maximum difference 3.6% in 11 year teacher rating).

The estimates for the regression and reliability coefficents are given in Table 3 on next page.

Percentage change between runs shows the following:

% change.	$\beta_{OLS}(11,16)$	$\beta_{IV}(11,16)$	$r(11)$	$\beta_{OLS}(7,11)$	$\beta_{IV}(7,11)$	$r(7)$
runs 1 v 1a	7.5	4.9	2.8	/	/	/
1a v 2	0.5	1.3	0.9	/	/	/
1 v 2	8.0	6.2	1.9	1.2	5.0	5.8
2 v 3-9	0.5	1.2	1.1	0.7	0.1	0.2

Here for the regression coefficients the differences between the runs using the different ancilliary variables (max. 1-2% for runs 2-9) are small in relation to the difference between "complete case" and "full use of ancilliary variables" datasets (1 v 2, max difference 8.0% in $\beta_{OLS}(11,16)$. However, the estimate of reliability at 11 years shows comparable variation (1.1% v 1.9%). In contrast to the means, these coefficients showed larger variation between the "complete case" and "key" data sets than between the key and corrected datasets, the latter difference being comparable to that between the corrected datasets. Notice that no information is available on the key datasets for the relation between 7 and 11

year tests and for the 7 year variables as the 7 year variables are not included in the set of key variables.

Table 3 shows that the runs excluding 16 year data give the same pattern to the estimates as those including 16 year data. However, systematic changes are found with inclusion of extra ancilliary variables with at one extreme the inclusion of only social class at 11 years (run 9) and of only social class at 11 years and ancilliary attainments at 11 years (run 8) producing no change from the "key" dataset in the β_{os} (11,16) and β_{jv}(11,16) respectively. Deletion of 1 of the ancilliary variables at each age, as in runs 3, 4, produced small changes in the regression and reliability coefficeints as did deletion of variables measured at age 16 (run 6).

Table 3 Effect on regression and reliability coefficents of estimation using different combinations of ancilliary variables

Run No.	β_{os}(11,16)	β_{jv}(11,16)	r(11)	β_{os}(7,11)	β_{jv}(7,11)	r(7)
1(complete case)	0.710	0.883	0.804	0.0590	0.0811	0.728
1a (key)	0.768	0.929	0.827	-	-	-
2 (all ancilliary)	0.772	0.941	0.820	0.0597	0.0772	0.723
3	0.774	0.940	0.823	0.0596	0.0771	0.773
4	0.770	0.930	0.829	0.0594	-	-
5	0.768	-	-	0.0593	-	-
6	0.772	0.936	0.824	0.0595	0.0772	0.771
7	0.773	0.936	0.826	0.0595	0.0772	0.771
8	0.771	0.929	0.829	0.0593	-	-
9	0.768	-	-	0.0593	-	-

CONCLUSIONS

The estimates of instrumental regression and reliability coeffficients given in Table 3 are obtained on a dataset which involved deletion of cases on which tests or teacher ratings on a number of attainments and social class had missing values at any age. These included 27.5% of the total cases, a similar percentage to the 32.1% of the "complete case" data above. The present analyses suggest that the "complete case" data shows biases in relation to the dataset deleting cases on all but 3 key variables, the "key" dataset, both in regression coefficients and reliability estimates and to a smaller extent in mean values. The

analysis of non response of Goldstein (in Fogelman(1976) is on different data than the "complete case" data used in some analyses in Goldstein (1979) and on analyses in Appendix B. Furthermore these analyses, in selecting non-responses on one test variable only, did not adequately reveal the effect of non-response. The interpolation procedure for missing values of Beale and Little allows an increase in the effective information utilised in the data. This gives further changes in the estimates usually in the same direction as between "complete case" and "key" datasets. These are large in comparison to the previous difference for the mean values and differ little when different sets of ancilliary variables are used. In contrast for the regression and reliability estimates further changes are of the same order as that between "complete case" and "key" datasets, that for the instrumental variable estimate of regression of 16 year on 11 score being 1.3% and for the reliability of 11 years score being 0.9%. In addition the use of different sets of ancilliary variables gave different estimates within roughly the same range. If the estimates obtained from the use of different subsets of ancilliary variables were reasonably constant we could be reasonably happy in concluding, given the range of variables used, that the total effect of "missingness" was largely taken up by these variables. As it is, it is possible that further controls using extra ancilliary variables will produce further changes in the coefficients, perhaps of the order of 1-2%.

The missing values interpolation program has been shown to be effective in utilising information on other relevant variables and the characteristics of the data are seen to effect the estimates, particularly of the "higher order" statistics such as regression and reliability estimates. Unlike many other interpolation techniques (see Marini, Olsen and Rubin, 1980) the Beale and Little technique produces unbiased estimates of these quantities when a certain assumption about the missing values given the observed data, the "missing at random" assumption, hold. The consistent change in coefficients as more and more data in the study is utilised is seen as evidence that the correction of missing values using these and possibly other variables should be a prerequisite for further

analyses with the data. Remaining problems for study are the calculation of the effective degrees of freedom of the corrected data, the sampling variability of the estimates produced and the utility of these estimates with data where the proportion of missing values is large (the present data had only 21% of the total data on the 10 variables missing).

APPENDIX 9

REGRESSION OF ATTAINMENT ON SCHOOL SOCIAL MIX WITHIN SOCIAL CLASS - EFFECT OF ERROR IN SOCIAL CLASS

By Russell Ecob

INTRODUCTION

Schooling tends to be competitive: in terms of measured attainment for example, some pupils succeed while others fail. It could be argued that our education system is geared to produce just this result, but much educational research seeks the social determinants of academic success as an aid to understanding inequalisty in society in broader terms, and also as a guide to educational policies that might reduce inequality.

The social class of a child's parents - usually measured by occupational categories by British studies - has long been acknowledged as one of the most important determinants of later academic success. The 'social mix' of a school has been seen as of additional importance, particularly because it is amenable to change as a policy tool - viz the interest in bussing of pupils and the advocacy of socially mixed schools as opposed to community or neighbourhood schooling:

> A long-established strategy (for redistribution of eductional resources) is through the socially mixed school where it is assumed that not only will children from all social backgrounds have the same access to resources but also, because of the presence of children who know how to demand, use and respond to resources effectively, those who would not otherwise do so fully will come to do so.
>
> (Eggleston, 1977, p 61)

In educational research, the framework for assessing the effect on pupil attainment is expressed by the linear model:

$$y = \alpha + \beta x + \varepsilon \qquad -(1)$$

Where y is pupil attainment, often a standardised test score;

α_i is the advantage to a pupil of social class i;

β_x is the advantage to all pupils in a school due to its social mix x; and

ε is the residual attainment, assumed random.

The estimated coefficients (α_i) represent an average social class effect within schools. The estimated β represents the effect of school social mix, net of any differences due to pupil social class. In practice, of course, the model is elaborated by the addition of further controls at both pupil and school levels, such as the pupil's sex, family size, ethnic origin and previous performance, and school size, type and location.

An early and influential report from the United States that examined this model found school social mix to be the most important single determinant of attainment: "Schools are remarkably similar in the way they relate to the achievement of their pupils when the socio-economic background of the students is taken into account" (Coleman et al, 1966, p 21). In Britain, Joan Barker Lunn (1971) found a similar result; the Inner London Education Authority's 'Literacy Survey' also found children of all social classes attaining better in the schools with more non-manual pupils in them, though the effect was greater for non-manual children than for working class children (Mabey, 1974); the Plowden Report on primary schooling (1967) advocated socially mixed schools as one method of assisting disadvantaged children. A summary and discussion of research on school social mix and pupil attainment is given by Simpson (1981)

Social class is prone to measurement error. Most of the studies mentioned have relied on the pupil's teacher for an assessment of their parental occupation, an assessment which is known not to be especially reliable.

This article is concerned with the effect that measurement error in social class categories will have on the coefficients of model (1) above. In particular, the estimated school social mix effect is shown to be considerably inflated by such measurement error.

UNIT OF AGGREGATION FOR SOCIAL MIX

The models we present are general in as much as the Social Mix is considered in relation to any aggregate of individuals. This can be either a classroom and year group or a whole school, and different choices may be more reasonable in different contexts. We will use the word 'group' in the following description. It should be noted that in the example used later, from the ILEA Literacy Survey, this denotes year group as information is only available on year groups, not on individual classrooms. In the examples given in the introduction the unit of aggregation was the school.

The unit of analysis in each case is the individual, who is conceived of as having three relevant items of information; attainment, social class and social mix of the group of which he/she is a member

A MODEL WITH TWO SOCIAL CLASSES

We consider first the simple model comprising two social classes (1,2) with the same proportion in the population. We denote the true social class by S_T and the observed social class by S_o and we suppose that the error of observation is such that the conditional probability of observing the wrong social class is ρ, independent of the true social class.

Thus $P(S_o=1 \mid S_T=2) = P(S_o=2 \mid S_T=1) = \rho$

We suppose also that the relation of attainment to Social Mix is linear within each true social class, having the same slope for each individual social class and that each group is of the same size (n). For simplicity we use the total number, R, in the group who are in social class 1 as the independent variable to represent the social mix.

We aim to express the observed slope in relation to the true slope and will expect the relationship to depend on three factors:

(a) difference of intercept of the true regression lines, assumed parallel;

(b) the distribution of social class mix within each true social class and their central tendencies;

(c) The conditional misclassification probability, p, which is assumed to be indpendent of the Social Class Mix (SCM), and of attainment.

Let us first make the additional assumption that there is no outside influence on class formation. We will later allow for outside influence.

If R = no of Social Class members in the group, then

$$P(R=r) = \binom{n}{r}\left(\frac{1}{2}\right)^n$$

Given R = r, P (true Social Class 1 member is observed) $= \frac{r}{n}$

Using the relationship, $P(S_T=1)\,P(R=r \mid S_T=1) = P(S_T=1 \mid R=r)\,P(R=r)$

we obtain $P(R=r \mid S_T=1) = \frac{r}{n}\binom{n}{r}\left(\frac{1}{2}\right)^n = \binom{n-1}{r-1}\left(\frac{1}{2}\right)^{n-1} \quad r=1,\ldots,n$

or, equivalently, $P(R=r+1 \mid S_T=1) = \binom{n-1}{r}\left(\frac{1}{2}\right)^{n-1} \quad r=0,\ldots,n-1$

Similarly $P(R=r \mid S_T=2) = \frac{n-r}{n}\binom{n}{r}\left(\frac{1}{2}\right)^n = \binom{n-1}{r}\left(\frac{1}{2}\right)^{n-1}, \quad r=0,1,\ldots,n-1$

Thus, the conditional distributions are identical, and binomial apart from a change in central tendency, the mean of the $S_T = 1$ distribution being $\frac{n+1}{2}$ and of the $S_T = 2$ distribution being $\frac{n-1}{2}$.

This analysis will later be adapted to allow for outside influences on class formation by supposing the two distributions are as follows

$$P(R=r+k \mid S_T=1) = \binom{n-k}{r}\left(\frac{1}{2}\right)^{n-k} \quad r=0,\ldots,n-k$$

$$P(R=r \mid S_T=2) = \binom{n-k}{r}\left(\frac{1}{2}\right)^{n-k} \quad r=0,\ldots,n-k$$

where $n > k \geq 1$. Useful values of $\frac{k}{n}$ may be estimated from the observed data making allowance for the effect of measurement error. The reasonableness of the binomial distributional assumption in this case will be tested on observed data.

Now let the equations for the true social classes be

$y = \alpha_1 + \beta_i x + e$ for Social Class 1

$y = \alpha_2 + \beta_i x + e$ for Social Class 2

where β_i : true regression coefficient

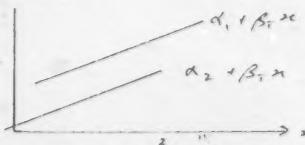
where y = attainment (possibly corrected for intake)

and x = no in Social Class 1

and where the errors are assumed to be identically and independently normally distributed.

The slope of the relation of attainment (y) to Social Mix (x) within observed social class 1 is obtained by fitting a straight line to the following situation:

attainment (y)



observed with probability $(1-\rho)$

observed with probability ρ

Then $\beta_{obs} = \dfrac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} x_{ij} y_{ij} P(x_{ij})}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} x_{ij}^2 P(x_{ij})}$

where i denotes social class and n_i is the number in the group in social class i,

where $P(x_{ij})$ is the overall probability of observing x_{ij} and

where x_{ij} is the deviation from the overall mean, μ.

where $\mu = (1-\rho)\left(\frac{n+1}{2}\right) + \rho\left(\frac{n-1}{2}\right)$

(or more generally, $\mu = (1-\rho)\left(\frac{n+k}{2}\right) + \rho\left(\frac{n-k}{2}\right)$)

if μ_1 is mean of social mix for Social Class 1, $\mu_1 - \mu = \rho\left[\frac{n+1}{2} - \frac{n-1}{2}\right] = \rho$

if μ_2 is mean of social mix for Social Class 2, $\mu_2 - \mu = 1-\rho$

(or more generally, $\mu_1 - \mu = k\rho$

$\mu_2 - \mu = k(1-\rho)$)

Substituting for y_{ij} and summing the separate components of the quadratic terms we obtain, after some algebra,

$$\beta_{obs} = \beta_T + \frac{\rho(1-\rho)(\alpha_1 - \alpha_2)}{\rho(1-\rho) + (n-1)/4} \qquad (1)$$

or, more generally, $\beta_{obs} = \beta_T + \frac{\rho(1-\rho)(\alpha_1 - \alpha_2)K}{\rho(1-\rho)K^2 + (n-K)/4}$ (2)

let j be defined by $\beta_T n = j(\alpha_1 - \alpha_2)$ then when j = 1 the effect of changing social class is the same as the maximum effect of social mix within social class. Then for various valued of $K = \frac{k}{n}$, j and p we calculate the multiplying fraction, f, where $\beta_{obs} = \beta_T * f$

where, approximately, $f = \left(1 + \frac{1}{jK}\right)$

where k = 1 we get

$$\beta_{obs} = \beta_T \left(1 + \frac{1}{j\left[K + (1-K)/4nK\rho(1-\rho)\right]}\right) \qquad (3)$$

and when k = 1, $f = 1 + \frac{1}{j\left[\frac{1}{n} + \frac{1+\frac{1}{n}}{4\rho(1-\rho)}\right]} \simeq \frac{1 + 4\rho(1-\rho)}{j}$

Values of f are given in Table 1 for different values of P_{ij} and K where n = 30. To obtain plausible values of ρ we use data on repeated observations of reported social class from the British Election Study and the Oxford Social Mobility Study. These studies are described in Appendix // and give respectively values of of 0.02 and 0.04*. It is not known how the size of error of measurement from pupils'reports compares with these but we would expect them to be at least as high.

TABLE 1: VALUES OF f FOR DIFFERENT VALUES OF ρ_{ij} AND K

K =			$0\left(\frac{1}{n}\right)$ (k=1)	0.1	0.2	0.5
p = 0.02	j	0.2	1.39	2.27	3.63	6.4
		0.5	1.16	1.57	2.05	3.16
		1	1.08	1.25	1.53	2.08
p = 0.05	j	0.2	1.95	4.28	7.0	8.59
		0.5	1.38	2.31	3.4	4.03
		1	1.19	1.66	2.20	2.52

EXTENSIONS OF THE MODEL TO SOCIAL CLASSES WITH DIFFERING PROPORTIONS

Let q = the overall proportion of true social class 1

The Conditional Binomial distributions now become

$$P\left(R=r+1 \mid S_i=1\right) = \binom{n-1}{r} q^r (1-q)^{n-1-r} = P(R=r \mid S_i=2)$$

This difference only enters in one term in equation (3) the value of f becoming

$$f = 1 + \rho(1-\rho)\left(j\left[\rho(1-\rho)K + q(1-q)(1-K)/nK\right]\right)^{-1}$$

There is little overall effect on f of differences in q in the range $.4 < q < 0.6$.

More extreme values within range $0.3 < q < 0.7$ have little effect also when k = 0.5 for the values of p in the range given though when k = 0.1 or less the term involving q(1-q) dominates the denominator. We illustrate in

the following calculations use the still higher value of 0.05

Table 2 the effect on f of varying q when there is no outside influence on class formation ($k=\frac{1}{n}$). The value of 0.2 is used for j this not effecting the relative influence of variation in q.

TABLE 2 VALUES OF f = β_{obs}/β_T FOR DIFFERENT VALUES OF q, ρ

	0.1	0.2	0.3	0.4	0.5	q
p = 0.02	2.12	1.63	1.48	1.42	1.40	
p = 0.05	3.68	2.52	2.16	2.02	1.95	

For more than two social classes the above analysis can be adapted, taking social classes in pairs and redefining the Social Mix to be the relative number in each class. The analysis only stricly applies however when the sum of individuals in the two social classes under consideration is constant in every group and it does not take account of the influence of social classes not included.

APPLICATION TO DATA FROM THE ILEA LITERARY SURVEY

Data on attainment of all 17564 pupils aged 8 in 503 ILEA schools in 1968 were used.

The value of $j = n\beta_T/(\alpha_1-\alpha_2)$

is not directly calculable as it depends on the unknown parameters $\beta_T, \alpha_1, \alpha_2$ of the true regression lines. However, β_{obs} and β_T are found to be related by the equation

$$\beta_T - \beta_{obs} = (\alpha_1-\alpha_2)_{obs}\ \frac{\rho(1-\rho)}{q(1-q)\left(\frac{n}{2}-1\right)(1-2\rho)} \qquad (4)$$

As n, the size of the year group is not constant, the average value is used. The social class division is that between manual and non-manual.

The following values of the parameters were found:

n = 43.21

k = 6.708

q = 0.234

$(\alpha_1 - \alpha_2)_{obs}$ = 6.867

β_{obs} = 0.338

From Equation 4 we obtain for, for $\rho = 0.02$ $\beta_T = \beta_{obs} - 0.143 = 0.195$

and for $\rho = 0.05$ $\beta_T = \beta_{obs} - 0.370 = -0.032$

So, using the conservative estimate of 0.02 for p a reduction of 42% of β from its former value is obtained, the influence of social mix changing sign between this and the higher estimate of measurement error.

REFERENCES

Barker Lunn, J (1971) Social Class, Attitudes and achievement. Slough, NFER.

Coleman J et al (1966) Equality of Educational Opportunity. Washington DC. US Department of Health, Education and Welfare.

Department of Education and Science (1967). Children and their primary schools. London, MHSO.

Eggleston, J (1977) Ecology of School. London, Methuen.

Mabey, C (1974) Social and ethnic mix and the relationship with attainment of children ages 8 and 11. London, Centre for Environmental Studies.

Simpson, L (1981) Statistical assessment of school effects using educational survey data. Doctoral thesis to be presented to the University of London.

APPENDIX 10

A DESCRIPTION OF TWO DATASETS USED TO ESTIMATE MEASUREMENT ERROR IN SOCIAL CLASS

By Russell Ecob

1 THE BRITISH ELECTION STUDY (BES)

Reinterview data gathered eight months apart was obtained from the British Election Study of the University of Essex, directed by Professor Bo Sarlvik and Ivor Crewe and conducted in February and October 1974, each interview following a General Election. Out of a sample of 1830 interviewed at both times, 1656 individuals were selected where the respondent was employed at both times. The analysis was on the 1097 reporting their own occupation eliminating those wives reporting their husband's occupation. The data was taken from an analysis by Fox and Alt (1976). Detailed job descriptions were obtained from the same person in both surveys and these were allocated to Occupational Unit Groups (OUG's). For the OUG's which were different at each occasion, a distinction was made between "genuine" and "spurious" change. The genuine changes in OUG are those believed to be caused by a genuine change in job, the spurious changes being those which, on examination of all occupation-related material, were believed to be caused by a description of the same job in a different way on each occasion. In addition, some changes in OUG are caused by coder error on either occasion. The reliability of social class coding is investigated on the subsample formed by eliminating those " genuine" OUG changes constituting 3.4% of the total sample which cause a change in social class.

2 THE OXFORD SOCIAL MOBILITY STUDY (OSMS)

The Oxford Social Mobility Study consists of a national survey in 1972 of 10309 men aged between 20 and 64, resident in England and Wales who were asked about their own and their father's education, their present occupation

and their father's occupation when they were 14. Two years later a reliability study was undertaken (Hope, Graham and Schwarz, 1979) which involved re-interviewing a representative 10% of the sample. The present data given in Table 2 comprises those 565 subjects who, when re-interviewed in 1974, maintained that their occupation was the same as that in 1972. This subsample has been shown to be representative of the complete sample. In terms of the six Registrar General's Classes, 28% of the subjects showed a change in Social Class in this period and for the aggregation into three classes given here, the figure was 10.3%.

The following breakdown into manual and non-manual social classes was found at each interview occasion for the two studies:

		nm	m			nm	m
BES	nm	553	19	OSMS	nm	164	28
	m	21	441		m	18	355

Assuming that the conditional misclassification probabilities (P) are identical for both classes, we obtain a value of 0.0195 for the BES study and a value of 0.042 for the OSMS study.